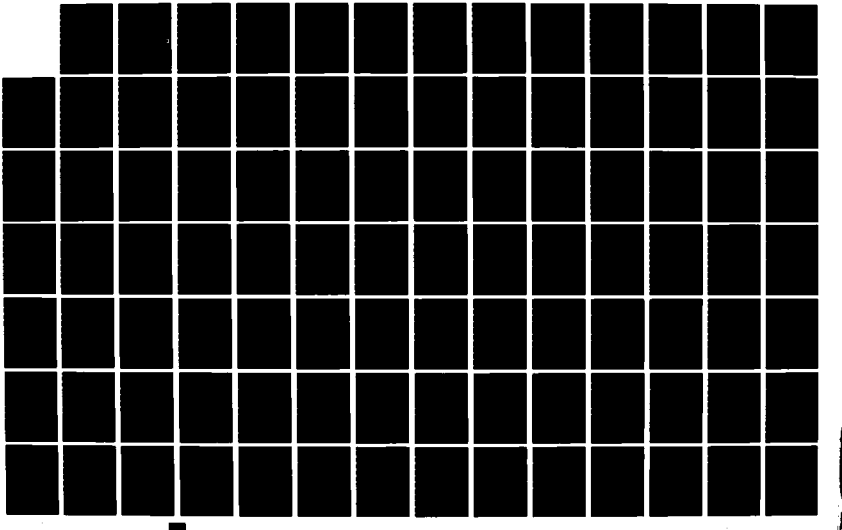
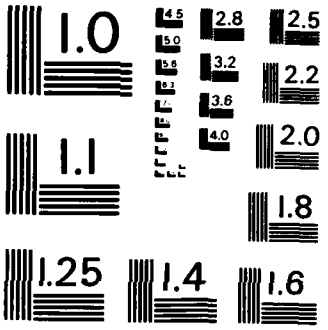
AD-A161 274 A MODEL FOR SERIAL DEPENDENCE IN LOGISTIC REGRESSION (U) MASSACHUSETTS INST OF TECH CAMBRIDGE STATISTICS CENTER T P LANE SEP 85 TR-38-ONR N00014-75-C-0555 1/2

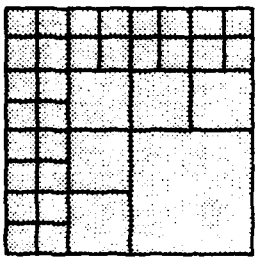
UNCLASSIFIED F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

(5)



STATISTICS CENTER

Massachusetts Institute of Technology

77 Massachusetts Avenue, Rm. E40-111, Cambridge, Massachusetts 02139 (617) 253-8722

A Model for Serial Dependence in Logistic Regression

by

Thomas Paul Lane

Massachusetts Institute of Technology

Technical Report No. ONR 38
Prepared under Contract
N00014-75-C-0555(NR-042-331)
for the Office of Naval Research

ABSTRACT

A model is proposed for binary time series with marginal probabilities given by logistic regression on explanatory variables, by analogy with the first order autoregressive error model for least squares regression. Measurements at adjacent time points are assumed to have an odds ratio that is not equal to one and that is constant as a function of time. Measurements separated in time are assumed to be conditionally independent given an intervening observation.

Consequences of using an ordinary logistic model in the presence of serial dependence are explored. The closest logistic model, defined as the one with the minimum Kullback-Leibler distance, is shown to be the one with the same marginal probabilities. Consistency of the maximum likelihood estimator of the serial dependence model is proved under certain conditions, and a procedure for finding these estimates is given.

Properties of the model are found, including expressions for the joint probabilities and the odds ratio between observations separated in time. The model is shown to generate *-mixing processes.

A score test is derived in order to test for independence after performing an ordinary logistic regression, and properties of this test are explored. The effects of missing data on the score test and on estimation of the odds ratio (with known coefficients) are presented.

The model is applied to the problem of automatic classification of EKG data based on feature extraction. A positive serial dependence is found in the examples presented.

A Model for Serial Dependence in Logistic Regression

by

Thomas Paul Lane

Massachusetts Institute of Technology

Technical Report No. ONR 38

Prepared under Contract

N00014-75-C-0555 (NR-042-331)

for the Office of Naval Research

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 The serial dependence model

Logistic regression is a common procedure for modeling a binary vector Y when there are explanatory variables. Under this model

$$\log \frac{P[Y_t=1]}{P[Y_t=0]} = X_t'\beta$$

where X is the vector of explanatory variables. This model assumes the observations on Y, given X, are independent. If Y is a time series, however, the independence assumption may not be realistic.

A similar problem can occur in ordinary least squares. If

$$Y_t = X_t'\beta + \varepsilon_t ,$$

it may be reasonable to assume the sequence $\{\varepsilon_t\}$ is serially correlated. The simplest model for a serially correlated series $\{\varepsilon_t\}$ is the first order autoregressive model, with

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

for some ρ less than 1 in absolute value and for a sequence of independent normally distributed $\{u_t\}$ with common mean 0 and unknown variance σ^2.

The parameter ρ measures the correlation between successive values of ε_t; this is a natural parameter to measure dependence between normal random variables. For binary variables, however, the odds ratio is in many re-

spects the natural parameter for measuring dependence.

This reasoning leads to the following model for serial dependence in logistic regression:

$$\log \frac{P[Y_t=1]}{P[Y_t=0]} = X_t'\beta \,, \qquad [1.1]$$

$$\frac{P[Y_t=Y_{t-1}=1]\ P[Y_t=Y_{t-1}=0]}{P[Y_t=1,Y_{t-1}=0]\ P[Y_t=0,Y_{t-1}=1]} = \psi \quad \text{for all } t, \qquad [1.2]$$

and Y_r and Y_t are conditionally independent given Y_s if $r<s<t$. (All probabilities are conditional on the $\{X_t\}$ sequence.) In the remainder of this paper I will refer to this model as the "serial dependence model."

1.2 Related models

Other models for binary time series have been proposed. For series without covariates, Billingsley (1961) considers stationary Markov processes. Keenan (1982) considers processes whose marginal probabilities are functions of an underlying stationary process. Kedem (1980) also considers stationary binary time series.

Logistic regression models for variables measured over time have been used in studies of "panel" or "longitudinal" data, in which repeated binary measurements are taken on a large number of subjects. Typically each individual time series is short, and any asymptotic theory that is developed holds as the number of subjects approaches infinity while the length of each series remains finite. Korn and Whittemore (1979) consider a model

in which the conditional probability of $\{Y_t=1\}$ is given by logistic regression on the covariates and on Y_{t-1}. (In models such as these Y_1 must be treated as a special case, and Korn and Whittemore assume the existence of $Y_0=Y_n$, where n is the length of the series.) Most other models also use a logistic function for the conditional probabilities.

Zeger et al. (1985) also consider longitudinal data, but their model is similar to the serial dependence model in that they model the marginal probabilities as logistic functions of the covariates. Apart from their application to longitudinal data, their model differs from the serial dependence model in two respects. First, they use the correlation between adjacent binary responses as their measure of association, and they assume it is constant. Second, their covariates are functions of the subject only and so do not vary with time.

The serial dependence model could be used for panel data, and many of the other models in the literature could be applied to a single long binary time series with time-varying covariates. However my motivation in proposing this model is the automatic EKG classification example in Chapter 8, so in this paper I will consider only a single long binary time series.

1.3 The odds ratio as a measure of association

Many measures of association are possible for binary random variables, but there are some desirable properties possessed only by those measures that are functions of the odds ratio. In this section I will discuss the possibility that other measures may be useful.

The analogy between the serial dependence model in logistic regression and the autoregressive error model in linear regression breaks down in the case of perfect association. In linear regression, if the correlation parameter is ± 1 and the coefficients are known, then knowledge of Y_t provides perfect knowledge of Y_s for all $s>t$, since $|\varepsilon_t|$ is then a constant for all t.

For binary variables an odds ratio of 0 or infinity does not provide an analogous property. Consider the following pair of two-by-two tables of joint probabilities of (A,B) and (C,D):

B \ A	1	0	
1	.6	.2	.8
0	.0	.2	.2
	.6	.4	

C \ D	1	0	
1	.6	.0	.6
0	.0	.4	.4
	.6	.4	

In both tables the odds ratio is infinite, but only for the pair (C,D) does knowledge of one member of the pair provide perfect knowledge of the other member. This happens only when the two random variables have the same marginal probabilities.

This feature of the odds ratio was observed by Feinberg (1981). He distinguishes between "complete" association, as in the (A,B) pair, and "absolute" association, as in the (C,D) pair. Since the serial dependence model does not distinguish between the two, it may not be a useful model in an application where it seems necessary for "perfect" association to imply "absolute" association. If correlation were used as the measure of association, this implication would hold. Correlation was used by Zeger et al. (1985).

Let the event $\{Y_{t-1} \neq Y_t\}$ be called a "state change." Examining the tables above shows that in the serial dependence model with an infinite odds ratio, state changes are possible if changing state would avoid moving to an event of smaller marginal probability. For example, let Y_{t-1} be A and let Y_t be B in the above table. The transition from {A=0} to {B=1} avoids the transition from {A=0} to {B=0}. Since P[A=0] > P[B=0], a state change is possible.

Data from a variety of sampling models can be entered in a two-by-two table. For example, fixed numbers of patients might be assigned to a "treatment" and a "control" group, and then might be classified as "improved" or "not improved" at the end of the study. One advantage of the odds ratio is that it is invariant to row and column multiplications, so in the hypothetical example it would not depend on the number of patients assigned to each group.

The above tables of marginal probabilities suggest a sampling model in which both classifications are random, and the row and column totals sum to 1. This is the case in the serial dependence model. In sampling models where the row and column totals are not arbitrary, the invariance property of the odds ratio loses some of its importance.

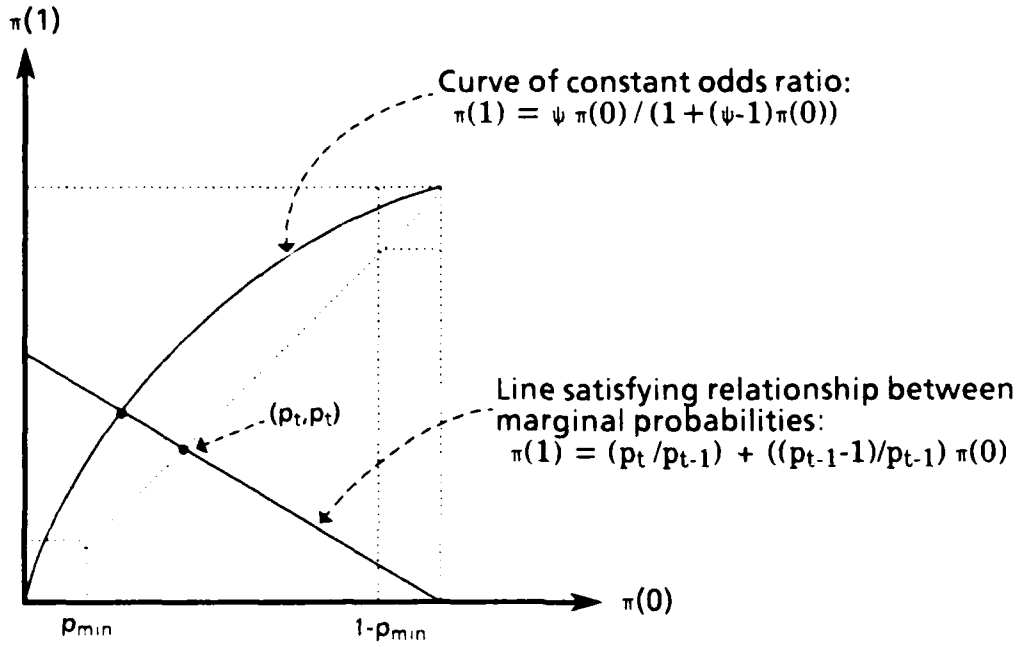
1.4 Notation

I will use the following notation in this paper. Relationships between some of these quantities are shown in Figure 1.1. Note that throughout

this paper the independent variables are treated as given, not as random variables.

Symbol	Meaning
Y_t	Dependent variable at time t
X_t	Vector of covariates at time t
p_t	Prob[Y_t=1] (marginal probability)
$\pi_t(j)$	Prob[Y_t=1|Y_{t-1}=j] (conditional probability)
α_t	Prob[Y_t=Y_{t-1}=1] (joint probability)
ψ	Odds ratio between successive observations
β	Vector of coefficients
$\hat{\psi}$, $\hat{\beta}$	Maximum likelihood estimates under the serial dependence model
$\tilde{\beta}$	Maximum likelihood estimate of β under the ordinary logistic model

Figure 1.1. Relationship Between Parameters of the Serial Dependence Model



For any given values of ψ, p_t, and p_{t-1}, the corresponding values of $\pi(1)$ and $\pi(0)$ are those at the intersection of the curve of constant odds ratio determined by ψ and the line determined by p_t and p_{t-1}. The quantity p_{min} is defined in Chapter 4.

Chapter 2

Properties of a Process Generated by the Model

In this chapter I will examine some properties of a process $\{Y_t\}$ generated by the serial dependence model. I will derive the joint distribution of two observations from the process and express the log linear representation of this joint distribution in terms of the quantities obtained from the model, namely the odds ratio and the marginal probabilities. I will derive an expression for the odds ratio between Y_t and Y_{t+n} for $n>1$. I will prove that $\{Y_t\}$ is a mixing process. Finally I will illustrate some of these properties with plots of the odds ratio between Y_t and Y_{t+n} as a function of n and φ.

2.1 The joint distribution of two observations

In this section I will obtain the joint distribution of two observations from a process generated by the serial dependence model, by relating the parameters of a log linear representation to the marginal probabilities and odds ratio. I will consider consecutive observations, but the same results apply to non-consecutive observations if the odds ratio between them is known. (A formula for the odds ratio between non-consecutive observations is given in the following section.)

For any given t let $p_t = P[Y_t=1]$ and $p_{t-1} = P[Y_{t-1}=1]$, and let φ be the (constant) odds ratio between Y_t and Y_{t-1}. Let $p_{ij} = P[Y_{t-1}=i, Y_t=j]$ be

decomposed as

$$\log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

with the usual constraints

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0.$$

First I will express the quantities obtained from the serial dependence model as functions of the parameters of the log linear representation. The marginal probabilities can be written as

$$\begin{aligned} p_t = p_{11}+p_{01} &= \exp(u+u_{1(1)}+u_{2(1)}+u_{12(11)})+\exp(u+u_{1(0)}+u_{2(1)}+u_{12(01)}) \\ &= \exp(u+u_{2(1)})\,[\exp(u_{1(1)}+u_{12(11)}) + \exp(u_{1(0)}+u_{12(01)})] \\ &= \exp(u+u_{2(1)})\,[\exp(u_{1(1)}+u_{12(11)}) + \exp(-u_{1(1)}-u_{12(11)})] \\ &= 2\exp(u+u_{2(1)})\cosh(u_{1(1)}+u_{12(11)}). \end{aligned}$$

Similarly

$$p_{t-1} = 2\exp(u+u_{1(1)})\cosh(u_{2(1)}+u_{12(11)}).$$

The odds ratio satisfies the equation

$$\begin{aligned} \log \psi &= \log p_{11} + \log p_{00} - \log p_{10} - \log p_{01} \\ &= (u+u_{1(1)}+u_{2(1)}+u_{12(11)}) + (u+u_{1(0)}+u_{2(0)}+u_{12(00)}) \\ &\quad - (u+u_{1(1)}+u_{2(0)}+u_{12(10)}) - (u+u_{1(0)}+u_{2(1)}+u_{12(01)}) \\ &= u_{12(11)}+u_{12(00)}-u_{12(10)}-u_{12(01)} = 4u_{12(11)}. \end{aligned}$$

The log linear parameters can be obtained by solving three equations, since given $u_{1(1)}$, $u_{2(1)}$, and $u_{12(11)}$, the other parameters are determined by the four constraints given above and the constraint $\sum_{ij} p_{ij} = 1$. The

two-factor term is, from above, $u_{12(11)} = (\log \varphi)/4$; it depends only on the odds ratio. Unfortunately the single-factor terms depend on the odds ratio and both marginal probabilities. They solve the pair of equations

$$p_t = 2 \exp(u+u_{2(1)}) \cosh (u_{1(1)}+(\log \varphi)/4)$$

$$p_{t-1} = 2 \exp(u+u_{1(1)}) \cosh (u_{2(1)}+(\log \varphi)/4) .$$

The parameter u can be removed by using these equations to obtain

$$\log \frac{p_t}{1-p_t} = 2u_{2(1)} + \log \cosh(u_{1(1)}+ \frac{1}{4} \log \varphi) - \log \cosh(u_{1(1)}- \frac{1}{4} \log \varphi)$$

and a similar equation for $\log [p_{t-1}/(1-p_{t-1})]$.

2.2 The joint distribution of three observations

For some $r<s<t$ let

$$p_{ijk} = P[Y_r=i, Y_s=j, Y_t=k]$$

be represented by a log linear model:

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} + u_{13(ik)} + u_{123(ijk)} , \qquad [2.1]$$

with the u-terms satisfying the constraints

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = \sum_j u_{23(jk)} = \sum_k u_{23(jk)} = \sum_i u_{13(ik)} = \sum_k u_{13(ik)} = \sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0. \qquad [2.2]$$

(These are not the same as the similarly named quantities in the previous section.) The u_{13} and u_{123} terms are 0 because Y_r and Y_t are independent given Y_s (see Fienberg, 1981, page 33).

Let ψ_{12} be the odds ratio between Y_r and Y_s and let ψ_{23} be the odds ratio between Y_s and Y_t. (If the Y's are consecutive observations, then $\psi_{12} = \psi_{23} = \psi$). The object of this section is to find ψ_{13}, the odds ratio between Y_r and Y_t.

Let a dot subscript denote summation over the corresponding index, so for example $p_{ij.} = \sum_k p_{ijk}$. This quantity can be written as

$$p_{ij.} = \exp(u+u_{1(i)}+u_{2(j)}+u_{3(1)}+u_{12(ij)}+u_{23(j1)})$$
$$+ \exp(u+u_{1(i)}+u_{2(j)}+u_{3(0)}+u_{12(ij)}+u_{23(j0)})$$
$$= 2 \exp(u+u_{1(i)}+u_{2(j)}+u_{12(ij)}) \cosh(u_{3(1)}+u_{23(j1)}) .$$

Summing over the other indices gives

$$p_{i.k} = 2 \exp(u+u_{1(i)}+u_{3(k)}) \cosh(u_{2(1)}+u_{12(i1)}+u_{23(1k)}) ,$$

$$p_{.jk} = 2 \exp(u+u_{2(j)}+u_{3(k)}+u_{23(jk)}) \cosh(u_{1(1)}+u_{12(1j)}) .$$

Then

$$\psi_{12} = (p_{11.}p_{00.})/(p_{10.}p_{01.}) ,$$

and

$$\log \psi_{12} =$$
$$\log[\exp(u+u_{1(1)}+u_{2(1)}+u_{12(11)})(\exp(u_{3(1)}+u_{23(11)})+\exp(u_{3(0)}+u_{23(10)}))$$
$$+\log[\exp(u+u_{1(0)}+u_{2(0)}+u_{12(00)})(\exp(u_{3(1)}+u_{23(01)})+\exp(u_{3(0)}+u_{23(00)}))$$
$$-\log[\exp(u+u_{1(1)}+u_{2(0)}+u_{12(10)})(\exp(u_{3(1)}+u_{23(01)})+\exp(u_{3(0)}+u_{23(00)}))$$
$$-\log[\exp(u+u_{1(0)}+u_{2(1)}+u_{12(01)})(\exp(u_{3(1)}+u_{23(11)})+\exp(u_{3(0)}+u_{23(10)}))$$

This can be simplified by using the constraints [2.2] to relate the

u-terms and obtain $u_{1(1)} = -u_{1(0)}$, etc. The result is

$$\log \psi_{12} = 4\, u_{12(11)} \,, \qquad [2.3]$$

or $u_{12(11)} = (\log \psi_{12})/4$. Similarly $u_{23(11)} = (\log \psi_{23})/4$.

The remaining odds ratio is given by

$$\log \psi_{13} =$$

$$u+u_{1(1)}+u_{3(1)}+\log[\exp(u_{2(1)}+u_{12(11)}+u_{23(11)})+\exp(u_{2(0)}+u_{12(10)}+u_{23(01)})]$$

$$+u+u_{1(1)}+u_{3(1)}+\log[\exp(u_{2(1)}+u_{12(01)}+u_{23(10)})+\exp(u_{2(0)}+u_{12(00)}+u_{23(00)})]$$

$$-u-u_{1(1)}-u_{3(1)}-\log[\exp(u_{2(1)}+u_{12(11)}+u_{23(10)})+\exp(u_{2(0)}+u_{12(10)}+u_{23(00)})]$$

$$-u-u_{1(1)}-u_{3(1)}-\log[\exp(u_{2(1)}+u_{12(01)}+u_{23(11)})+\exp(u_{2(0)}+u_{12(00)}+u_{23(01)})]$$

$$= \log[2\cosh(u_{2(1)}+u_{12(11)}+u_{23(11)})]+\log[2\cosh(u_{2(1)}-u_{12(11)}-u_{23(11)})]$$

$$-\log[2\cosh(u_{2(1)}+u_{12(11)}-u_{23(11)})]+\log[2\cosh(u_{2(1)}-u_{12(11)}+u_{23(11)})]$$

$$= \log \frac{\cosh[u_{2(1)}+(\log \psi_{12}+\log \psi_{23})/4]\,\cosh[u_{2(1)}-(\log \psi_{12}+\log \psi_{23})/4]}{\cosh[u_{2(1)}+(\log \psi_{12}-\log \psi_{23})/4]\,\cosh[u_{2(1)}+(\log \psi_{12}-\log \psi_{23})/4]}$$

This yields the relation

$$\psi_{13} = \frac{\cosh 2u_{2(1)} + \cosh[\,(\log \psi_{12}+\log \psi_{23})/2\,]}{\cosh 2u_{2(1)} + \cosh[\,(\log \psi_{12}-\log \psi_{23})/2\,]} \,. \qquad [2.4]$$

For the special case of three consecutive observations, $\psi_{12} = \psi_{23} = \psi$, and

$$\varphi_{13} = \frac{\cosh 2u_{2(1)} + \cosh \log \varphi}{\cosh 2u_{2(1)} + 1} \quad . \qquad [2.5]$$

These expressions can be used to calculate the dependence between observations separated in time. Suppose $Y_1,\ldots,Y_n$ have marginal probabilities $p_1,\ldots,p_n$ and common odds ratio φ betwen Y_t and Y_{t+1} for all t. Then $\varphi_{12}=\varphi$, and for any $t>2$, φ_{1t} can be found recursively by applying the above expression with marginal probabilities (p_1,p_{t-1},p_t) and odds ratios $\varphi_{1,t-1}$ and $\varphi_{t-1,t}=\varphi$.

Use of these formulas requires calculation of $u_{2(1)}$. The log linear representation provides the relation

$$u_{2(1)} = (\log p_{111} + \log p_{010} - \log p_{101} - \log p_{000})/4 \ .$$

The Markov property implies $p_{ijk} = P[Y_r=i,Y_s=j]\ P[Y_t=k|Y_s=j] = P[Y_r=i,Y_s=j]\ P[Y_s=j,Y_t=k] \ / \ p_s$. The pairwise joint probabilities can be obtained from the marginal probabilities and $\alpha=P[Y_r=i,Y_s=j]$. Equation [1.1] provides an expression for α:

$$\varphi_{12} = \frac{\alpha(1+\alpha-p_r-p_s)}{(p_r-\alpha)(p_s-\alpha)} \quad . \qquad [2.6]$$

This determines α uniquely, because the quadratic

$$\varphi(p_r-\alpha)(p_s-\alpha) - \alpha(1+\alpha-p_r-p_s) = 0$$

has only one root in the interval of acceptable α values, or

$$\max(p_r,p_s) < \alpha < \min(1,p_r+p_s).$$

This can be seen by examining Figure 1.1. It can be proved by noting that the right-hand-side of [2.6] increases from 0 at $\alpha=\max(p_r,p_s)$ to infinity at $\alpha=\min(1,p_r+p_s)$. It therefore takes the value φ_{12} an odd number of

times. A quadratic can have no more than two roots, so there is a single root in this interval.

2.3 A mixing property

There are various mixing properties that determine how dependence in a stochastic process dies off as a function of time. The strongest is *-mixing.

Using the expressions for the odds ratio between observations separated in time it is possible to bound the dependence between distant observations. Specifically, I will show that a process generated by the serial dependence model is a *-mixing process. I will use this property later to establish consistency of the maximum likelihood estimator.

Definition: A process is defined as a *-mixing process (Hall and Heyde, 1980, page 40) if there exist a number N and a function f defined on the positive integers such that

(1) f(n) is non-increasing in n for n>N;

(2) f(n) approaches 0 as n approaches infinity;

(3) for all t, given any event A in the σ-field generated by $\{Y_1,\ldots,Y_t\}$ and any event B in the σ-field generated by $\{Y_{t+n},\ldots\}$,

$$|P(AB) - P(A)\,P(B)| \leq f(n)\,P(A)\,P(B),$$

where AB is the intersection of A and B. By the Markov property, it is sufficient to consider only $A=\{Y_t=1\}$ or $\{Y_t=0\}$ and $B=\{Y_{t+n}=1\}$ or $\{Y_{t+n}=0\}$.

Take any fixed t and let φ_n be the odds ratio between Y_t and Y_{t+n}. Let $A=\{Y_t=1\}$ and $B=\{Y_{t+n}=1\}$. Then

$$\varphi_n = \frac{P(AB)[1+P(AB)-P(A)-P(B)]}{[P(A)-P(AB)][P(B)-P(AB)]},$$

$$\varphi_n - 1 = \frac{P(AB)-P(A)P(B)]}{[P(A)-P(AB)][P(B)-P(AB)]},$$

$$\begin{aligned} |P(AB)-P(A)P(B)| &= |\varphi_n-1|\ [P(A)-P(AB)]\ [P(B)-P(AB)] \\ &\leq |\varphi_n-1|\ P(A)\ P(B). \end{aligned}$$

The other combinations of A and B can be treated similarly, and two of these combinations lead to the bound $|(1/\varphi_n)-1|$. Therefore

$$|P(AB)-P(A)P(B)| \leq \max\{|\varphi_n-1|, |(1/\varphi_n)-1|\}\ P(A)\ P(B),$$

so it remains to show that φ_n approaches 1 uniformly in t. I will show this for $n=2^j$ by induction on j. Then I will extend the result to other values of n by showing $|\varphi_{n+1}-1| \leq |\varphi_n-1|$ for all n.

Suppose $\varphi>1$ (the proof for $\varphi<1$ is similar). The expressions given above for φ_{13} show that if $\varphi>1$, then $\varphi_n>1$ for all n. Therefore 0 is a lower bound on (φ_n-1) for all n in the following proof.

For each t there is a quantity u (equal to $u_{2(1)}$ in the log linear expansion carried out above) such that

$$\varphi_2 = \frac{\cosh 2u + \cosh \log \varphi}{\cosh 2u + 1}.$$

It is easy to see that if a, b, and c are positive and if $b>c$, then $(a+b)/(a+c)$ is a decreasing function of a; its derivative with respect to a is $(b-c)/(a+c)^2$. Since $\cosh x \geq 1$ for all x, it therefore follows that

for all t,

$$\varphi_2 \leq \frac{1 + \cosh \log \varphi}{2} = \frac{1}{2} + \frac{1}{4}\left(\varphi + \frac{1}{\varphi}\right) \leq \frac{3}{4} + \frac{\varphi}{4},$$

and therefore $(\varphi_2 - 1) \leq (\varphi - 1)/4$.

Now for a fixed positive integer j, let $n = 2^{j+1}$ and $k = 2^{-2j}$, and suppose for all t

$$(\varphi_{n/2} - 1) \leq k(\varphi - 1).$$

Fix t and let φ_A and φ_B be the odds ratios for the pairs $(Y_t, Y_{t+n/2})$ and $(Y_{t+n/2}, Y_{t+n})$, respectively. Then there is a value u such that

$$\varphi_n = \frac{\cosh 2u + \cosh\left[\frac{1}{2}\log \varphi_A + \frac{1}{2}\log \varphi_B\right]}{\cosh 2u + \cosh\left[\frac{1}{2}\log \varphi_A - \frac{1}{2}\log \varphi_B\right]}.$$

Then for all t,

$$\begin{aligned}
\varphi_n &\leq \frac{1}{2} + \frac{1}{2}\cosh\left[\frac{1}{2}\log\left(1+k(\varphi-1)\right) + \frac{1}{2}\log\left(1+k(\varphi-1)\right)\right] \\
&= \frac{1}{2} + \frac{1}{4}\left[\left(1+k(\varphi-1)\right) + \left(1+k(\varphi-1)\right)^{-1}\right] \\
&\leq \frac{3}{4} + \frac{1}{4}\left(1+k(\varphi-1)\right),
\end{aligned}$$

which implies $\varphi_n - 1 \leq k(\varphi-1)/4 = (\varphi-1)n^{-2}$. Therefore by induction, if f(n) is defined as $(\varphi-1)n^{-2}$, then $(\varphi_n - 1) \leq f(n)$ when n is a power of 2.

It remains to show that f can be suitably defined when n is not a power of 2. The above argument shows $\liminf (\varphi_n - 1) = 0$, so it is sufficient to show that $|\varphi_{n+1} - 1| < |\varphi_n - 1|$. But for any given t there is a quantity u such that

$$\varphi_{n+1} = \frac{\cosh 2u + \cosh\left[\frac{1}{2}\log \varphi_n + \frac{1}{2}\log \varphi\right]}{\cosh 2u + \cosh\left[\frac{1}{2}\log \varphi_n - \frac{1}{2}\log \varphi\right]}.$$

Then for all t,

$$\varphi_{n+1} \leq \frac{1 + \cosh\left[\frac{1}{2}\log\varphi_n + \frac{1}{2}\log\varphi\right]}{1 + \cosh\left[\frac{1}{2}\log\varphi_n - \frac{1}{2}\log\varphi\right]} = B_{n+1},$$

say. If $\varphi=1$, then $B_{n+1}=1$, and as φ approaches infinity, B_{n+1} approaches φ_n. This implies $\varphi_{n+1} \leq \varphi_n$, because B_{n+1} has no local maximum for $1\leq\varphi\leq\infty$, since its derivative is always positive:

$$\frac{\partial B_{n+1}}{\partial\varphi} = \frac{\sinh\left[\frac{1}{2}\log\varphi + \frac{1}{2}\log\varphi_n\right]}{2\varphi\left\{1+\cosh\left[\frac{1}{2}\log\varphi - \frac{1}{2}\log\varphi_n\right]\right\}}$$

$$- \frac{\sinh\left[\frac{1}{2}\log\varphi + \frac{1}{2}\log\varphi_n\right]\left\{1+\cosh\left[\frac{1}{2}\log\varphi + \frac{1}{2}\log\varphi_n\right]\right\}}{2\varphi\left\{1+\cosh\left[\frac{1}{2}\log\varphi - \frac{1}{2}\log\varphi_n\right]\right\}^2}$$

$$= \frac{1}{2\varphi}\left[\frac{\frac{1}{2}\left((\varphi\varphi_n)^{1/2}-(\varphi\varphi_n)^{-1/2}\right)}{1 + \frac{1}{2}\left((\varphi/\varphi_n)^{1/2}-(\varphi_n/\varphi)^{1/2}\right)}\right.$$

$$\left. - \frac{\frac{1}{2}\left((\varphi/\varphi_n)^{1/2}-(\varphi_n/\varphi)^{1/2}\right)\left[1 + \frac{1}{2}\left((\varphi\varphi_n)^{1/2}+(\varphi\varphi_n)^{-1/2}\right)\right]}{\left[1 + \frac{1}{2}\left((\varphi/\varphi_n)^{1/2}+(\varphi_n/\varphi)^{1/2}\right)\right]^2}\right].$$

Multiplying by the positive quantity

$$4\varphi\left[1 + \frac{1}{2}\left((\varphi/\varphi_n)^{1/2}+(\varphi_n/\varphi)^{1/2}\right)\right]^2$$

gives

$$((\varphi\varphi_n)^{1/2}-(\varphi\varphi_n)^{-1/2})\ \left[\ 1 + \frac{1}{2}\ ((\varphi/\varphi_n)^{1/2}+(\varphi_n/\varphi)^{1/2})\ \right]$$

$$- ((\varphi/\varphi_n)^{1/2}-(\varphi_n/\varphi)^{1/2})\ \left[\ 1 + \frac{1}{2}\ ((\varphi\varphi_n)^{1/2}+(\varphi\varphi_n))^{-1/2}\ \right]$$

$$= (\varphi\varphi_n)^{1/2}-(\varphi\varphi_n)^{-1/2}-(\varphi/\varphi_n)^{1/2}+(\varphi_n/\varphi)^{1/2} + \varphi_n - 1/\varphi_n$$

$$= (\varphi^{1/2}+\varphi^{-1/2})\ (\varphi_n^{1/2}-\varphi_n^{-1/2}) \ + \ (\varphi_n - 1/\varphi_n)\ ,$$

which is strictly positive if $\varphi_n>1$. (If $\varphi_n=1$, $\varphi_{n+1}=1$.)

Therefore if f is defined as

$$f(n) \ = \ \begin{cases} (\varphi-1)/n^2 & \text{if } n \text{ is a power of 2,} \\ f(n-1) & \text{otherwise,} \end{cases} \qquad [2.7]$$

then f satisfies the three conditions in the definition of a *-mixing sequence. Repeating this proof for the case $\varphi<1$ gives the following proposition.

Proposition: The serial dependence model generates *-mixing sequences.

2.4 Some numerical calculations

Figures 2.1 through 2.6 show the log of the odds ratio between Y_t and Y_{t+n} for $1\leq n\leq 10$, calculated for some special cases using the formulas derived in this chapter. The step functions are obtained from the upper bound f on (φ_n-1) derived in the previous section, equation [2.7]. For each curve the marginal probability p_t takes the constant value 0.5 for all t.

For plots with $\psi<1$, f is obtained by applying [2.7] to the process $\{Z_t\} = \{Y_1, 1-Y_2, Y_3, 1-Y_4, \ldots\}$; each even-numbered term is changed. If the original process has odds ratio ψ_{st} between Y_s and Y_t, the odds ratio betwen Z_s and Z_t is ψ_{st} if s-t is even and $1/\psi_{st}$ if s-t is odd. Therefore the upper bound obtained from f, and a lower bound that is the inverse of this upper bound, bound the odds ratio of the original process.

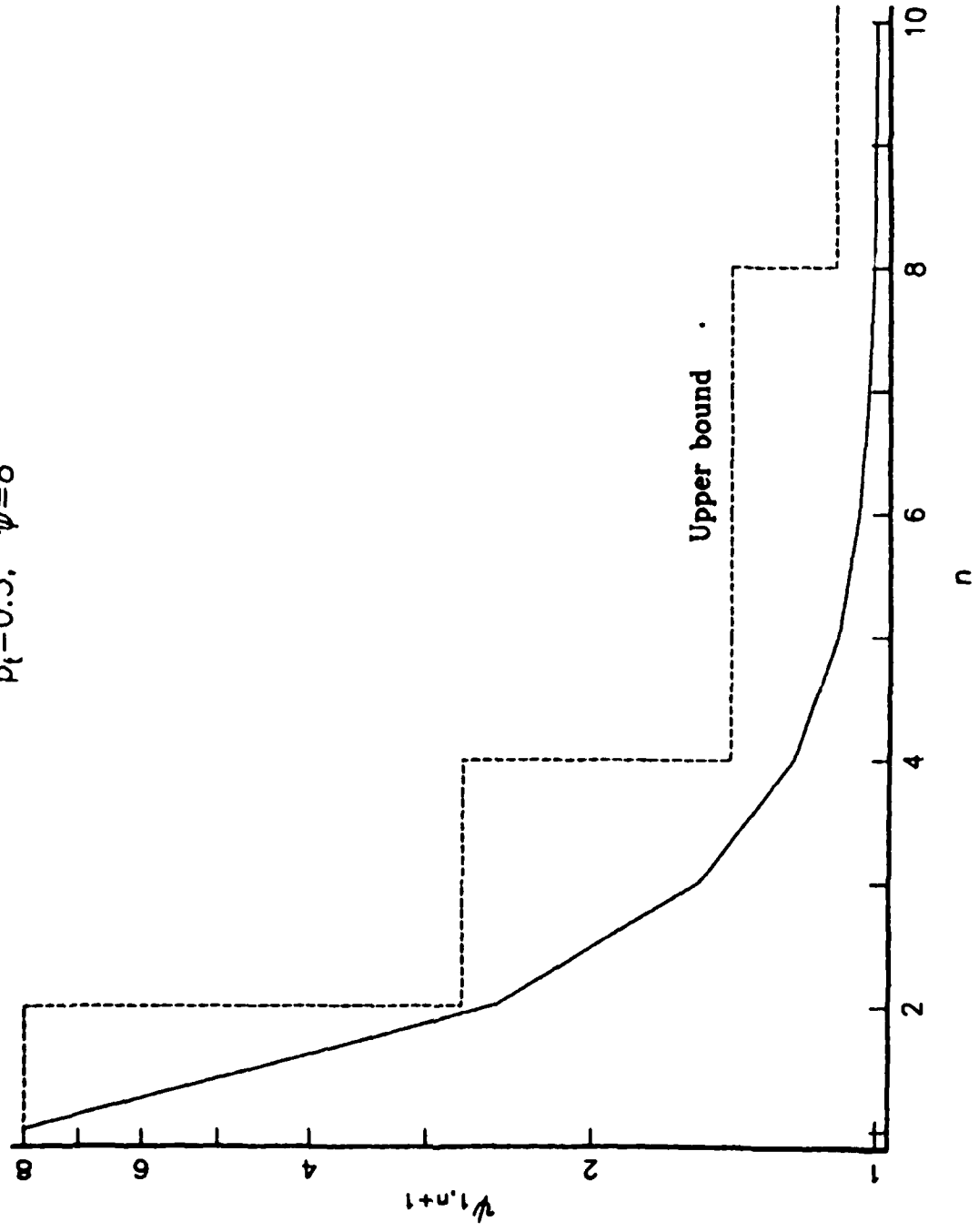


Figure 2.1. $\psi_{1,n+1}$ as a Function of n
p_t=0.5, ψ=8

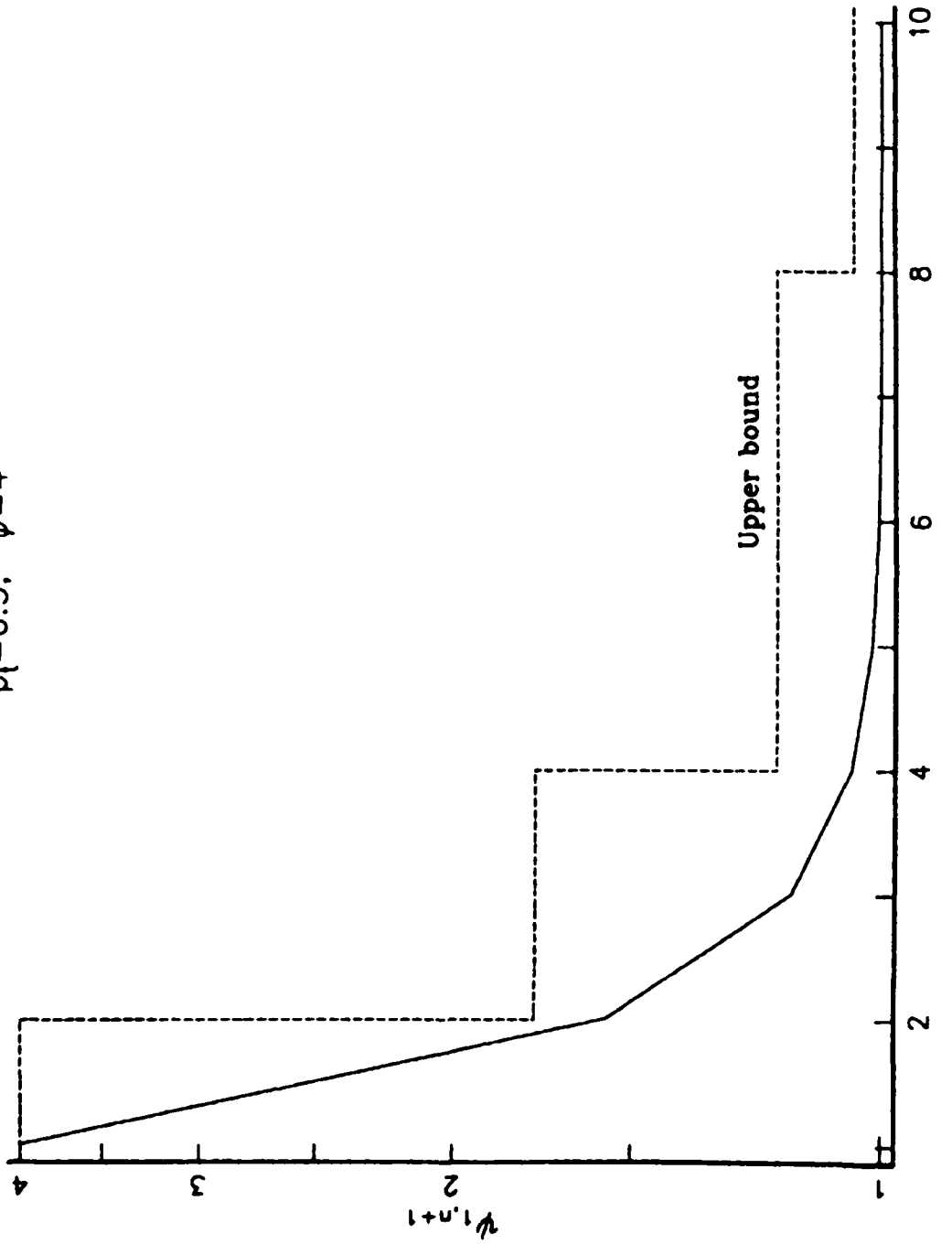


Figure 2.2. $\psi_{1,n+1}$ as a Function of n
p_t=0.5, ψ=4

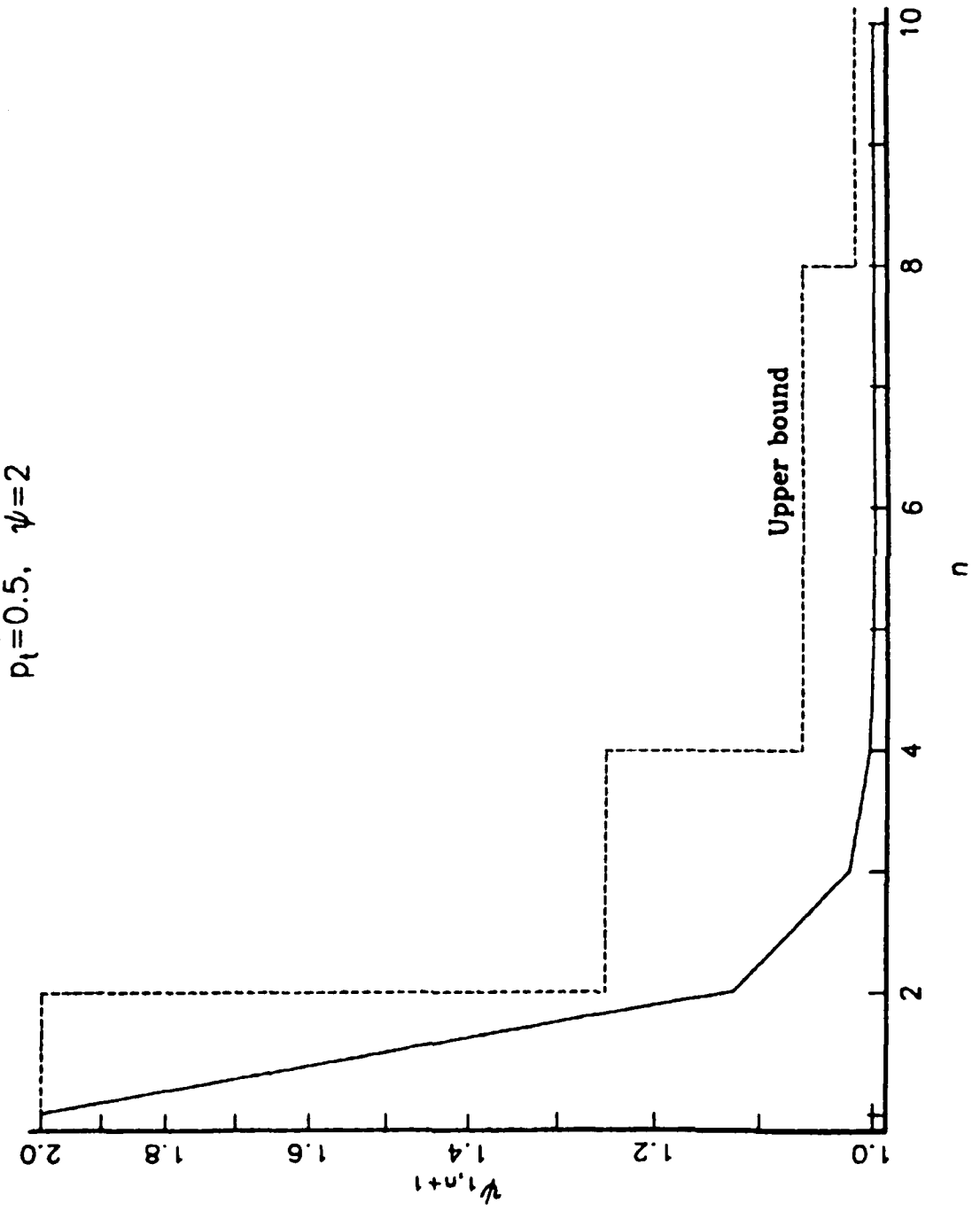


Figure 2.3. $\psi_{1,n+1}$ as a Function of n
$p_t=0.5$, $\psi=2$

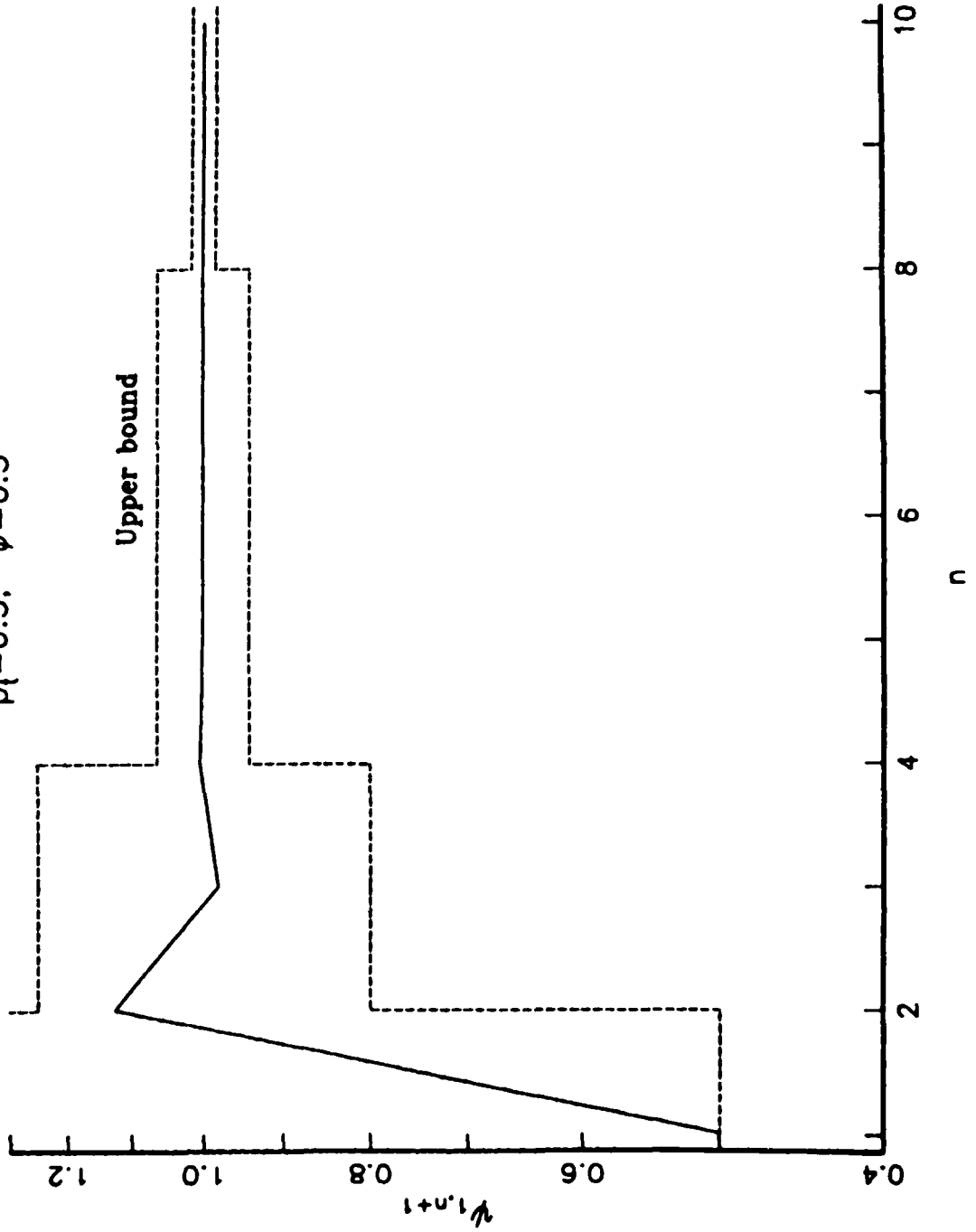


Figure 2.4. $\psi_{1,n+1}$ as a Function of n
p_i=0.5, ψ=0.5

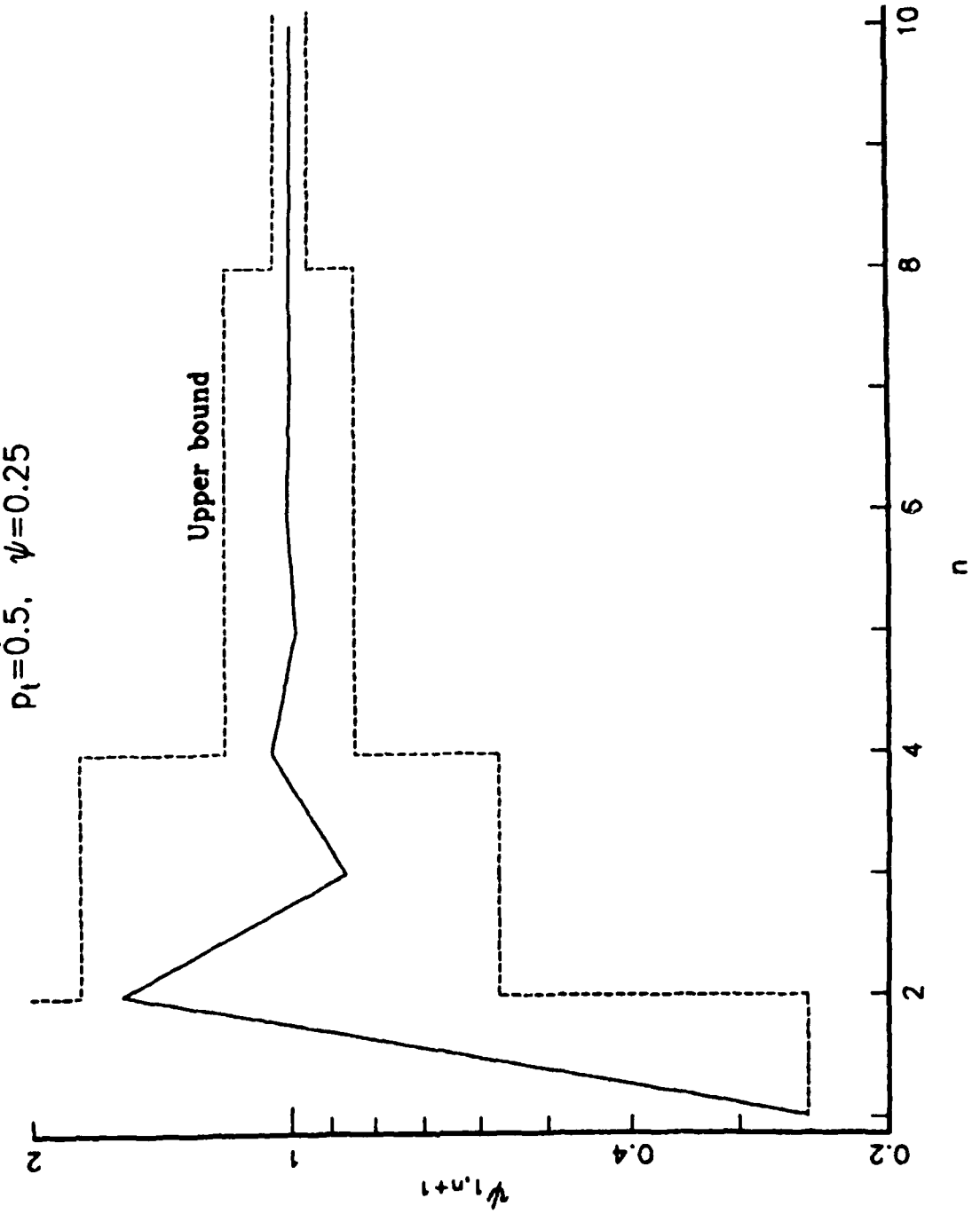


Figure 2.5. $\psi_{1,n+1}$ as a Function of n
p_1=0.5, ψ=0.25

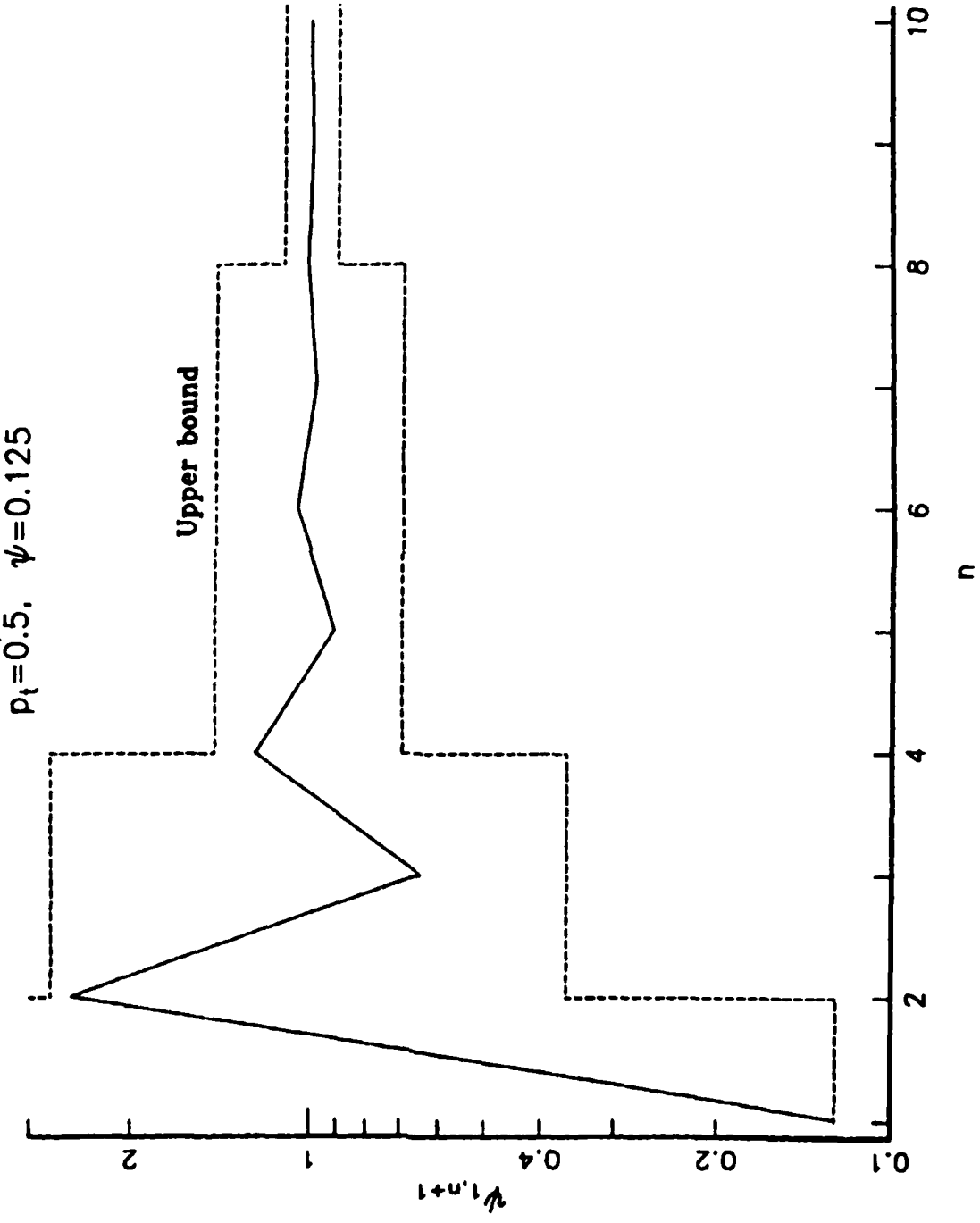


Figure 2.6. $\psi_{1,n+1}$ as a Function of n
p_t=0.5, ψ=0.125

Chapter 3

Consequences of Ignoring Serial Dependence

In linear regression with serially correlated errors, coefficients estimated by ordinary least squares are consistent, but the estimated standard errors are not correct. In this chapter I will show that the same holds for the serial dependence model. I will give some theoretical indication that the ordinary logistic coefficient estimates are consistent estimators of the coefficients in the serial dependence model, and I will verify this with a simulation. I will also perform another simulation in which I will show that confidence intervals computed using the standard errors from the logistic model do not have the correct coverage probabilities.

3.1 Coefficient estimates

Suppose a process is generated by the serial dependence model with unknown coefficients β_0. I will show that the nearest ordinary logistic model to the serial dependence model, in the sense of minimum Kullback-Leibler distance, is the one with the same coefficients. Since these coefficients maximize the expected values (under the model that generated the process) of log of the ordinary logistic likelihood, this is an indication that the ordinary logistic coefficient estimates should converge to the true values.

Let f be the density function in the true model, and let $\{p_t\}$ and $\{\alpha_t\}$ be

the usual marginal and joint probabilities in that model. Let g be the density function for an ordinary logistic model, and let $\{q_t\}$ be the marginal probabilities in that model. They can be written

$$\log g = \log P[Y_1=y_1]+\log P[Y_2=y_2|Y_1=y_1]+\ldots+\log P[Y_n=y_n|Y_{n-1}=y_{n-1}]$$
$$= \Sigma_t \, [y_t \log q_t + (1-y_t) \log (1-q_t)]$$

and

$$\log f = y_1 \log p_1 + (1-y_1) \log (1-p_1) + \Sigma_{t\geq 2} \, y_{t-1}y_t \log (\alpha_t/p_{t-1})$$
$$+ (1-y_{t-1})y_t \log((p_t-\alpha_t)/(1-p_{t-1})) + y_{t-1}(1-y_t) \log(p_{t-1}-\alpha_t)/p_{t-1}$$
$$+ (1-y_{t-1})(1-y_t) \log ((1+\alpha_t-p_t-p_{t-1})/(1-p_{t-1})) \, .$$

Then the Kullback-Leibler distance is

$$K_{fg} = E_f \, [\log (f/g) \,]$$
$$= p_1 \log \frac{p_1}{q_1} + (1-p_1) \log \frac{1-p_1}{1-q_1} + \sum_{t=2}^{n} \alpha_t \log \frac{\alpha_t}{p_{t-1}q_t}$$
$$+ (p_t-\alpha_t) \log \frac{p_t-\alpha_t}{(1-p_{t-1}) \, q_t} + (p_{t-1}-\alpha_t) \log \frac{p_{t-1}-\alpha_t}{p_{t-1}(1-q_t)}$$
$$+ (1+\alpha_t-p_{t-1}-p_t) \log \frac{1+\alpha_t-p_{t-1}-p_t}{(1-p_{t-1})(1-q_t)} \qquad [3.1]$$

To minimize this distance it is helpful to collect terms and write

$$K_{fg} = A - p_1 \log q_1 - (1-p_1) \log (1-q_1) - \{\Sigma_{t\geq 2} \, \alpha_t \log q_t$$
$$+ (p_t-\alpha_t)\log q_t + (p_{t-1}-\alpha_t)\log (1-q_t) + (1+\alpha_t-p_{t-1}-p_t)\log (1-q_t)\}$$
$$= A-p_1\log q_1-(1-p_1)\log(1-q_1)-\{\Sigma_{t\geq 2} \, p_t\log q_t+(1-p_t)\log(1-q_t)\} \, ,$$

where A is a function of $\{p_t\}$ and $\{\alpha_t\}$ but not $\{q_t\}$. The derivative with respect to q_t is

$$\frac{\partial K_{fg}}{\partial q_t} = -p_t \frac{1}{q_t} + (1-p_t) \frac{1}{1-q_t} ,$$

which is 0 if and only if $p_t = q_t$. The second derivative is positive for all values of p_t and q_t between 0 and 1, so the Kullback-Leibler distance is a minimum if $q_t=p_t$ for all t. This condition is satisfied if the coefficients in the two models are the same, so the following proposition is proved.

Proposition: The closest ordinary logistic model to any serial dependence model is the one with the same coefficients, if distance is measured by the Kullback-Leibler distance using the serial dependence model as the true model.

Let $\lambda = \log \varphi$, and suppose the coefficients in the two models are the same. The value of the Kullback-Leibler distance between the two models is well approximated by a fourth order Taylor series around $\lambda=0$. If $\lambda=0$, then $K_{fg}=0$ because the two models coincide. From the proof of the proposition, $\partial K_{fg}/\partial\lambda = 0$ at $\lambda=0$. Taking derivatives in [3.1] and collecting terms with logarithms gives

$$\frac{\partial K_{fg}}{\partial \varphi} = \sum_{t=2}^{n} \frac{\partial a_t}{\partial \varphi} \log \frac{a_t(1+a_t-p_t-p_{t-1})}{(p_t-a_t)(p_{t-1}-a_t)} + \frac{\partial a_t}{\partial \varphi} (1 - 1 - 1 + 1)$$

$$= \sum_{t=2}^{n} \frac{\partial a_t}{\partial \varphi} \log \varphi = \lambda \sum_{t=2}^{n} \frac{\partial a_t}{\partial \varphi} .$$

Applying the chain rule and taking repeated derivatives lead to

$$\frac{\partial K_{fg}}{\partial \lambda} = \lambda e^{\lambda} \sum \frac{\partial a_t}{\partial \varphi}$$

$$\frac{\partial^2 K_{fg}}{\partial\lambda^2} = (\lambda+1)e^{\lambda}\sum\frac{\partial a_t}{\partial\varphi} + \lambda e^{2\lambda}\sum\frac{\partial^2 a_t}{\partial\varphi^2}$$

$$\frac{\partial^3 K_{fg}}{\partial\lambda^3} = (\lambda+2)e^{\lambda}\sum\frac{\partial a_t}{\partial\varphi} + (3\lambda+2)e^{2\lambda}\sum\frac{\partial^2 a_t}{\partial\varphi^2} + \lambda e^{3\lambda}\sum\frac{\partial^3 a_t}{\partial\varphi^3}$$

$$\frac{\partial^4 K_{fg}}{\partial\lambda^4} = (\lambda+3)e^{\lambda}\sum\frac{\partial a_t}{\partial\varphi} + (7\lambda+9)e^{2\lambda}\sum\frac{\partial^2 a_t}{\partial\varphi^2} + (6\lambda+3)e^{3\lambda}\sum\frac{\partial^3 a_t}{\partial\varphi^3} + \lambda e^{4\lambda}\sum\frac{\partial^4 a_t}{\partial\varphi^4}$$

Equation [1.2] or [2.6] defines α in terms of φ and the marginal probabilities. The first derivative of α_t with respect to φ is

$$\frac{\partial a_t}{\partial\varphi} = \frac{(p_t-a_t)(p_{t-1}-a_t)}{1+(\varphi-1)(p_t-p_{t-1}-2a_t)} . \qquad [3.2]$$

Higher derivatives can be obtained from this expression. At $\lambda=0$, the derivatives simplify to

$$\partial\alpha/\partial\varphi = p_t(1-p_t)\ p_{t-1}(1-p_{t-1})$$

$$\partial^2\alpha/\partial\varphi^2 = -2\ (\partial\alpha/\partial\varphi)\ [p_t(1-p_{t-1}) + p_{t-1}(1-p_t)]$$

$$\partial^3\alpha/\partial\varphi^3 = 6(\partial\alpha/\partial\varphi)\ [p_t^2(1-p_{t-1})^2+p_{t-1}^2(1-p_t)^2+3p_t(1-p_t)p_{t-1}(1-p_{t-1})] .$$

Therefore the Taylor series up to λ^4 gives

$$K_{fg}(\lambda) \approx (1/2)\ \lambda^2 \sum p_t(1-p_t)p_{t-1}(1-p_{t-1})$$

$$+ (1/3)\ \lambda^3 \sum p_t(1-p_t)(1-2p_t)p_{t-1}(1-p_{t-1})(1-2p_{t-1})$$

$$+ (1/8)\ \lambda^4 \sum p_t(1-p_t)p_{t-1}(1-p_{t-1})\ [1-6p_t(1-p_{t-1})-6p_{t-1}(1-p_t)$$

$$+6p_t^2(1-p_{t-1})^2+6p_{t-1}^2(1-p_t)^2+18p_t(1-p_t)p_{t-1}(1-p_{t-1})]\ ,$$

so it follows that

$$K_{fg}(\lambda) = (1/2)\ \lambda^2 \sum p_t(1-p_t)p_{t-1}(1-p_{t-1})$$

$$[1+(2/3)\lambda(1-2p_t)(1-2p_{t-1})+O(\lambda^2)]. \qquad [3.3]$$

This shows that given a sequence of $\{p_t\}$ such that $\{p_t-1/2\}$ and $\{p_{t-1}-1/2\}$ tend to have the same sign, the logistic model is closer to models with negative λ than to those with positive λ. The converse is true if $\{p_t-1/2\}$ tends to oscillate in sign.

Figure 3.1 shows the exact Kullback-Leibler distance between the two models for one example. Here I generated 100 independent normal random variables $\{X_t\}$ and used an intercept and slope both equal to 1.0. I calculated the Kullback-Leibler distance for various values of the odds ratio equally spaced on the log scale between 0.1 and 10. The curves are values given by the first one, two, and three non-zero terms in the Taylor series in λ. The series up to λ^4 gives a good fit to the exact distances except for very low values of the odds ratio, where the two models are more distant than the approximation suggests.

To examine the coefficient estimates using the logistic model with data generated under the serial dependence model, I performed a simulation. The conditions were identical to those described in the previous para-

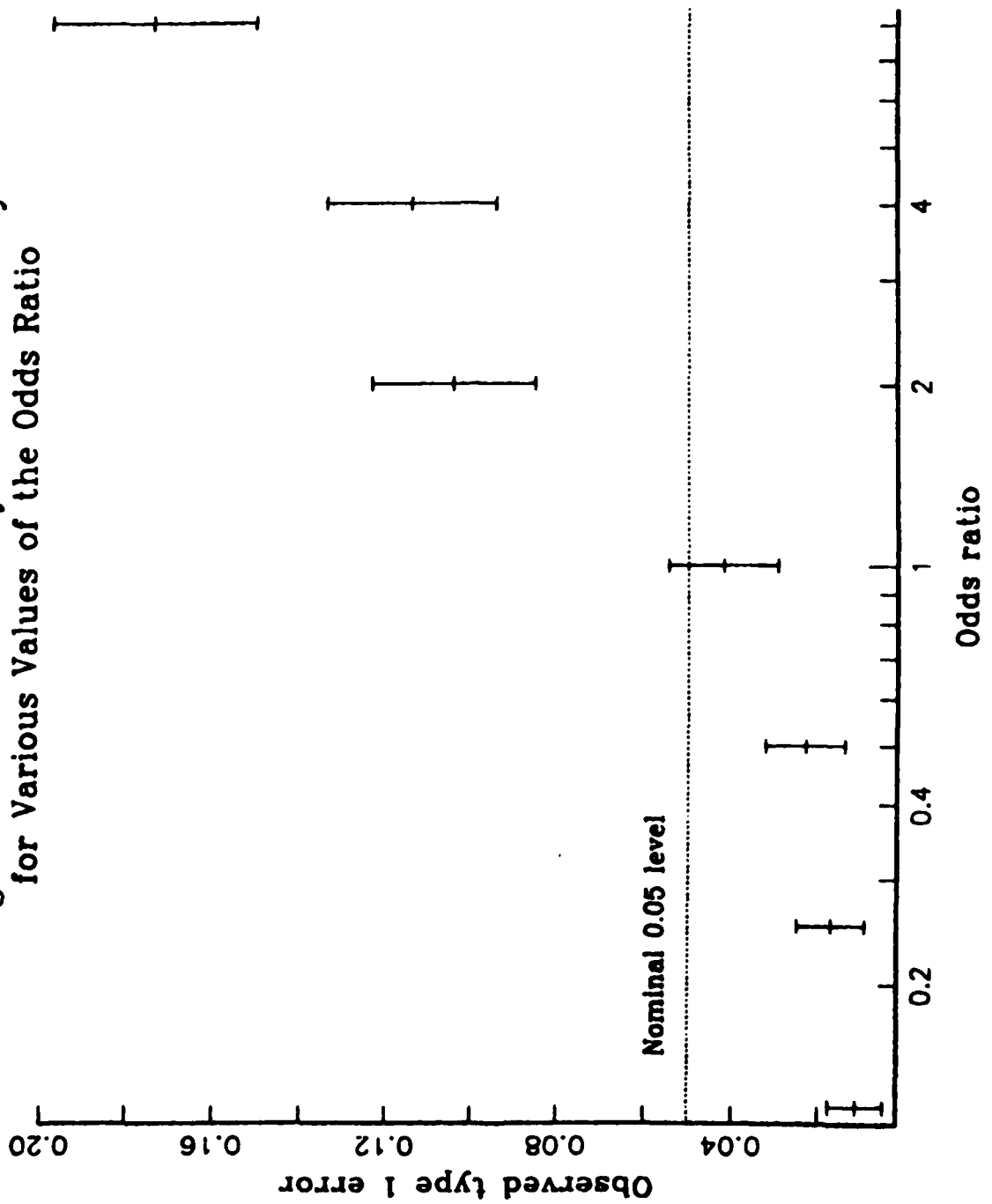


Figure 3.1. Observed Rejection Probability for Various Values of the Odds Ratio

graph. The results are shown in Table 3.1.

It appears that each of the quantities in the table is an increasing function of the odds ratio. The quantities associated with the intercept, however, increase much faster. The sample bias and standard deviation both double as the odds ratio moves from 0.1 to 10. For the slope, the bias increases by a substantial proportion but the standard deviation is more nearly constant.

The bias is too small a fraction of its estimated standard error to confirm the effect with a hypothesis test. For example, the difference between the bias at ψ=10 and that at ψ=0.1 is about equal to its estimated standard error. However the obvious trend indicates that the observed difference is not an artifact of the simulation.

3.2 Standard Errors

The ordinary logistic coefficient estimates differ from the true coefficient values only by a small fraction of their standard deviations. However Table 3.1 shows that the standard deviations of the intercept estimates change by a factor of two as the odds ratio changes from 0.1 to 10. This is an indication that the standard errors produced by the logistic model cannot be correct. Therefore inferences about the coefficients are suspect.

This fact is more apparent in Figure 3.2. This shows the result of a second simulation, with 1000 observations at each value of the odds ratio

Table 3.1. Ordinary Logistic Estimates

Observed bias and standard deviation from a sample of 400 ordinary logistic estimates for each value of the odds ratio. Each regression has 100 observations and a single explanatory variable. The true intercept and slope values are both 1.0.

Odds	----- Intercept ----		----- Slope ----	
Ratio	Bias	Std Dev	Bias	Std Dev
0.10	.025	.203	.048	.308
0.13	.026	.208	.054	.302
0.16	.029	.214	.061	.297
0.20	.024	.215	.063	.296
0.25	.024	.216	.060	.296
0.32	.023	.223	.061	.296
0.40	.027	.229	.057	.296
0.50	.031	.241	.061	.304
0.63	.031	.242	.060	.304
0.79	.034	.247	.060	.305
1.00	.038	.256	.061	.311
1.26	.040	.272	.065	.312
1.58	.040	.281	.066	.321
2.00	.035	.294	.064	.328
2.51	.044	.309	.066	.327
3.16	.044	.322	.073	.340
3.98	.042	.338	.079	.335
5.01	.042	.352	.083	.339
6.31	.047	.375	.084	.340
7.94	.053	.391	.088	.350
10.00	.054	.402	.097	.357

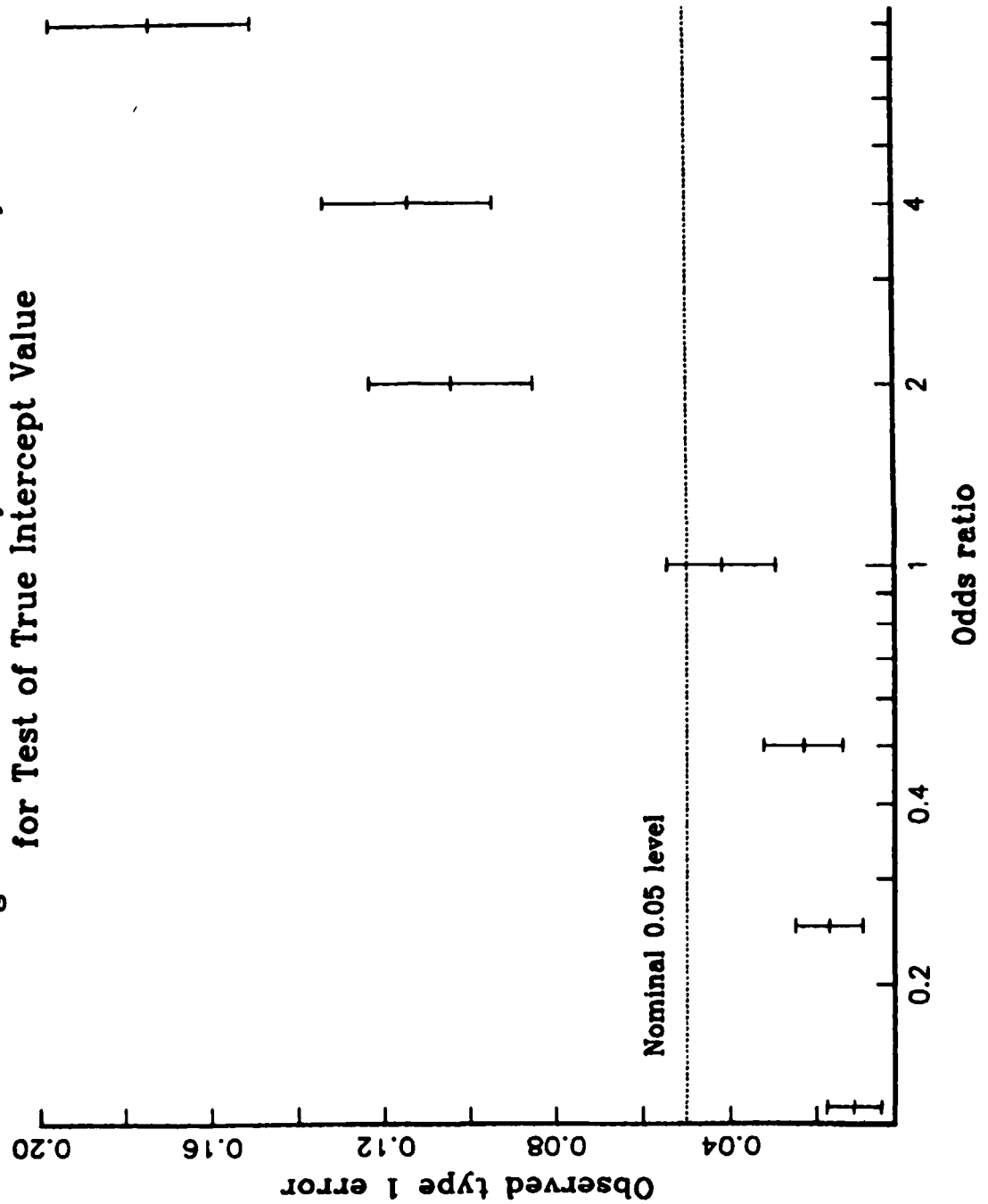


Figure 3.2. Observed Rejection Probability for Test of True Intercept Value

but otherwise with conditions identical to those in the previous simulation. In each sample I counted the number of "misses," or the number of times the true value of the intercept was not within 1.96 standard errors of the estimated value, with both estimates and standard errors from the ordinary logistic model. This is a test of size 0.05 using the asymptotic normal distribution of the estimates.

For each sample the proportion $\hat{\theta}$ of misses is an estimate of the size θ of the test. Since the number of misses follows a binomial distribution, a 95 percent confidence interval for the size is $\hat{\theta} \pm 1.96(\hat{\theta}(1-\hat{\theta})/1000)^{1/2}$. The Figure shows this confidence interval for each sample, along with a line marking the nominal 0.05 level.

Clearly this test does not have the proper size if the odds ratio is not equal to 1. For smaller odds ratios the estimated standard errors in the logistic model are too large, so the intercept estimates are more accurate than they appear. For larger odds ratios the situation is worse. Confidence intervals computed using the ordinary logistic model are too optimistic; their coverage probabilties are much smaller than their nominal values.

Chapter 4

Maximum Likelihood Estimation

In this Chapter I write the likelihood function and its derivatives. I give conditions that imply consistency of the maximum likelihood estimates, and I describe a method for finding these estimates. I also compare the maximum likelihood estimator with some other estimators.

4.1 Solution of the likelihood equations

If the likelihood equation is written as the product of conditional likelihoods, its logarithm can be written

$$L = y_1 \log p_1 + (1-y_1) \log(1-p_1) + \sum_{t=2}^{n} y_t y_{t-1} \log \frac{a_t}{p_{t-1}}$$

$$+ y_t(1-y_{t-1}) \log \frac{p_t - a_t}{1-p_{t-1}} + (1-y_t)y_{t-1} \log \frac{p_{t-1} - a_t}{p_{t-1}}$$

$$+ (1-y_t)(1-y_{t-1}) \log \frac{1+a_t-p_t-p_{t-1}}{1-p_{t-1}}, \qquad [4.1]$$

where $a_t = \text{Prob}[Y_t=Y_{t-1}=1]$. This quantity is defined by equation [2.6], and its derivatives are

$$\frac{\partial a}{\partial \varphi} = \frac{(p_t-a_t)(p_{t-1}-a_t)}{1-(\varphi-1)(2a_t-p_t-p_{t-1})}$$

$$\frac{\partial a}{\partial \beta} = \frac{p_t(1-p_t)x_t[\varphi p_{t-1}-(\varphi-1)a_t] + p_{t-1}(1-p_{t-1})x_{t-1}[\varphi p_t-(\varphi-1)a_t]}{1-(\varphi-1)(2a_t-p_t-p_{t-1})}$$

The derivatives of the log likelihood are most easily expressed in terms

of the derivatives of α and the marginal probabilities. The marginal probability p_t does not depend on φ but its derivative with respect to β is $p_t(1-p_t)x_t$.

The derivative of the log likelihood with respect to φ is

$$\frac{\partial L}{\partial \varphi} = \sum_{t=2}^{n} \frac{\partial \alpha_t}{\partial \varphi} \left[\frac{y_t y_{t-1}}{\alpha_t} - \frac{y_t(1-y_{t-1})}{p_t - \alpha_t} - \frac{y_{t-1}(1-y_t)}{p_{t-1} - \alpha_t} + \frac{(1-y_t)(1-y_{t-1})}{1+\alpha_t - p_t - p_{t-1}} \right].$$

The expectation of the term in brackets is 0, so the second derivative of α does not enter into the Fisher information matrix. Terms of that matrix that involve derivatives with respect to φ are

$$- E \left[\frac{\partial^2 L}{\partial \varphi^2} \right] = \sum_{t=2}^{n} \left[\frac{\partial \alpha_t}{\partial \varphi} \right]^2 \left[\frac{1}{\alpha_t} + \frac{1}{p_t - \alpha_t} + \frac{1}{p_{t-1} - \alpha_t} + \frac{1}{1+\alpha_t - p_t - p_{t-1}} \right] \quad [4.2]$$

$$- E \left[\frac{\partial^2 L}{\partial \beta \partial \varphi} \right] = \sum_{t=2}^{n} \frac{\partial \alpha_t}{\partial \varphi} \frac{\partial \alpha_t}{\partial \beta} \left[\frac{1}{\alpha_t} + \frac{1}{p_t - \alpha_t} + \frac{1}{p_{t-1} - \alpha_t} + \frac{1}{1+\alpha_t - p_t - p_{t-1}} \right]$$

$$- \left[\frac{\partial \alpha_t}{\partial \varphi} \right] \left[\frac{p_t(1-p_t)(1-p_{t-1})x_t}{(p_t - \alpha_t)(1+\alpha_t - p_t - p_{t-1})} + \frac{p_{t-1}(1-p_{t-1})(1-p_t)x_{t-1}}{(p_{t-1} - \alpha_t)(1+\alpha_t - p_t - p_{t-1})} \right]. \quad [4.3]$$

The remaining submatrix requires the derivative of the log likelihood with respect to β:

$$\frac{\partial L}{\partial \beta} = \sum_{t=2}^{n} \frac{\partial \alpha_t}{\partial \beta} \left[\frac{y_t y_{t-1}}{\alpha_t} - \frac{y_t(1-y_{t-1})}{p_t-\alpha_t} - \frac{y_{t-1}(1-y_t)}{p_{t-1}-\alpha_t} + \frac{(1-y_t)(1-y_{t-1})}{1+\alpha_t-p_t-p_{t-1}} \right]$$

$$+ p_t(1-p_t)x_t \left[\frac{y_t(1-y_{t-1})}{p_t-\alpha_t} - \frac{(1-y_t)(1-y_{t-1})}{1+\alpha_t-p_t-p_{t-1}} \right]$$

$$+ p_{t-1}(1-p_{t-1})x_{t-1} \left[\frac{y_{t-1}(1-y_t)}{p_{t-1}-\alpha_t} - \frac{(1-y_t)(1-y_{t-1})}{1+\alpha_t-p_t-p_{t-1}} \right]$$

$$- \sum_{t=2}^{n-1} p_t(1-p_t)x_t \ .$$

Here again the terms in brackets have zero expectation, so this simplifies the expresssion for the information matrix somewhat.

$$- E\left[\frac{\partial^2 L}{\partial \beta^2} \right] = \sum_{t=2}^{n} \frac{\partial \alpha_t}{\partial \beta} \frac{\partial \alpha_t'}{\partial \beta} \left[\frac{1}{\alpha_t} + \frac{1}{p_t-\alpha_t} + \frac{1}{p_{t-1}-\alpha_t} + \frac{1}{1+\alpha_t-p_t-p_{t-1}} \right]$$

$$- \frac{\partial \alpha_t}{\partial \beta} \left[\frac{p_t(1-p_t)(1-p_{t-1})x_t}{(p_t-\alpha_t)(1+\alpha_t-p_t-p_{t-1})} + \frac{p_{t-1}(1-p_{t-1})(1-p_t)x_{t-1}}{(p_{t-1}-\alpha_t)(1+\alpha_t-p_t-p_{t-1})} \right]'$$

$$- \left[\frac{p_t(1-p_t)(1-p_{t-1})x_t}{(p_t-\alpha_t)(1+\alpha_t-p_t-p_{t-1})} + \frac{p_{t-1}(1-p_{t-1})(1-p_t)x_{t-1}}{(p_{t-1}-\alpha_t)(1+\alpha_t-p_t-p_{t-1})} \right] \frac{\partial \alpha_t'}{\partial \beta}$$

$$+ \frac{p_t^2(1-p_t)^2(1-p_{t-1})x_t x_t'}{(p_t-\alpha_t)(1+\alpha_t-p_t-p_{t-1})} + \frac{p_{t-1}^2(1-p_{t-1})^2(1-p_t)x_{t-1}x_{t-1}'}{(p_{t-1}-\alpha_t)(1+\alpha_t-p_t-p_{t-1})}$$

$$+ \frac{p_t(1-p_t)p_{t-1}(1-p_{t-1})}{1+\alpha_t-p_t-p_{t-1}} (x_t x_{t-1}' + x_{t-1}x_t')$$

$$- \sum_{t=2}^{n-1} p_t(1-p_t)\, x_t x_t' \ . \qquad [4.4]$$

4.2 Consistency of the maximum likelihood estimates

Most proofs of the consistency and asymptotic normality of maximum likeli-

hood estimators involve assumptions about and manipulation of the Fisher information matrix. This is difficult here because the information is unwieldy. However a proof of consistency is possible under some conditions.

Wald (1949) proved the consistency of the maximum likelihood estimator for independent and identically distributed random variables under certain regularity conditions. A simplified version was given by Chernoff (1972). These proofs did not make use of the derivatives of the likelihood function. A similar proof can be used here by noting that the log of the likelihood ratio (of the likelihood at an alternative parameter value to the likelihood at the true parameter value) is a supermartingale, and by applying a strong law of large numbers.

The strong law requires some degree of independence. As was proved in Chapter 2, if $\{Y_t\}$ is generated by the serial dependence model then it is a *-mixing sequence. It is easy to see that if $\{Y_t\}$ is also a Markov process, then $f(Y_t,\ldots,Y_{t+s})$ is a *-mixing sequence for any function f and integer s. Strong laws are available for such sequences.

To prove consistency it is necessary to be able to distinguish the <u>true</u> parameter value $\theta_0=(\log \varphi_0,\beta_0)$ from any alternative value $\theta=(\log \varphi,\beta)$. (Values below calculated at $\theta=\theta_0$ are given the subscript 0.) To do so it is necessary that the conditional probabilities $\underline{\pi}_{t0} = (\pi_{t0}(0),\pi_{t0}(1)) = (P[Y_t=1|Y_{t-1}=0,\theta_0],P[Y_t=1|Y_{t-1}=1,\theta_0])$ be occasionally different from the probabilities $\underline{\pi}_t = (\pi_t(0),\pi_t(1))$ computed under the alternative θ. Reference to Figure 1.1 indicates that these conditional probabilities coincide

if $\psi=\psi_0$ and if $(p_{t-1,0}, p_{t0}, p_{t-1}, p_t)$, where

$$\log [p_t/(1-p_t)] = X_t'\beta ,$$

falls on a set $S_p=S_p(\theta_0,\theta)$ for which

$$p_{t-1,0}\ \pi_{t0}(1) + (1-p_{t-1,0})\ \pi_{t0}(0) = p_{t0} \qquad [4.5]$$

$$p_{t-1}\ \pi_{t0}(1) + (1-p_{t-1})\ \pi_{t0}(0) = p_t \qquad [4.6]$$

$$\frac{\pi_{t0}(1)(1-\pi_{t0}(0))}{(1-\pi_{t0}(1))\pi_{t0}(0)} = \psi_0 \qquad [4.7]$$

for some $\underline{\pi}_{t0}$.

The set S_p corresponds to a set $S_x=S_x(\theta_0,\theta)$ in the (X_{t-1},X_t) space. As long as a substantial proportion of the pairs (X_{t-1},X_t) are removed from S_x, one may expect to be able to discriminate between θ_0 and θ.

Consistency follows from the following assumptions.

[A1] Bounded covariates: there exists a positive M such that for all t, $|X_t| \leq M$.

This assumption implies the existence of a positive p^* such that for all t, $p^* < \min(p_{t0}, 1-p_{t0})$. This in turn implies the existence of p_{min} such that $0 < p_{min} \leq p^*$ and for all t, $p_{min} < \min(\pi_{t0}(1), 1-\pi_{t0}(1), \pi_{t0}(0), 1-\pi_{t0}(0))$.

[A2] Identifiability: given $\beta \neq \beta_0$, there exist $\varepsilon=\varepsilon(\beta_0,\beta)>0$, $\eta=\eta(\beta_0,\beta)$, and $T=T(\beta_0,\beta)$ such that

$$\eta < n^{-1} \#\{t: 2 \leq t \leq n \text{ and } d((X_{t-1},X_t),S_x) \geq \varepsilon\}, \qquad [4.8]$$

where

#A is the number of elements in the set A,

d(z,S) is the minimum distance between z and S, and

$S_x = S_x(\theta_0, \theta)$, with $\theta = (\log \varphi_0, \beta)$.

This assumption states that a non-negligible proportion of the (X_{t-1}, X_t) pairs do not lie arbitrarily close to the manifold on which $\underline{\pi}_t = \underline{\pi}_{t0}$. It appears awkward, but following the proof I will expand upon this condition and give other conditions that imply it.

[A3] Compact parameter space: the true values of the parameters can be assumed to lie in a known compact region.

With this assumption we can consider the restricted maximum likelihood estimator subject to membership in the compact set.

The first two assumptions imply the following two lemmas. The first shows that a non-negligible proportion of the $\underline{\pi}_t$'s are bounded away from $\underline{\pi}_{t0}$. The second shows that a non-negligible proportion of the log likelihood ratios are bounded below zero.

Lemma 1. If Assumptions [A1] and [A2] are satisfied and if $\theta = (\log \varphi_0, \beta)$, $\beta \neq \beta_0$, then there exist $\varepsilon_1 = \varepsilon_1(\theta_0, \theta)$, $\eta = \eta(\theta_0, \theta)$, and $T = T(\theta_0, \theta)$ such that for all $n > T$,

$$\eta < n^{-1} \#\{t: 2 \leq t \leq n \text{ and } d(\underline{\pi}_t, \underline{\pi}_{t0}) \geq \varepsilon\}. \qquad [4.9]$$

Proof: The set S_x is the set of (X_{t-1}, X_t) for which $\underline{\pi}_t = \underline{\pi}_{t0}$. Let

$$D_\varepsilon = \{(X_{t-1},X_t):\ |X_{t-1}|\leq M,\ |X_t|\leq M,\ d((X_{t-1},X_t),S_t)\geq\varepsilon\}. \quad [4.10]$$

Since $d(\underline{\pi}_t,\underline{\pi}_{t0})$ is a continuous function that is positive on the compact set D_ε, it achieves some positive minimum value $\varepsilon_1(\varepsilon,\theta_0,\theta)$ on it. Assumption [A2] therefore establishes the lemma.

Lemma 2. If Assumption [A1] and the conclusion of Lemma 1 hold, there exist $\eta=\eta(\theta_0,\theta)>0$, $\varepsilon_2=\varepsilon_2(\theta_0,\theta)$ and $T=T(\theta_0,\theta)$ such that for all $n>T$,

$$\eta < n^{-1}\ \#\{t:\ 2\leq t\leq n \text{ and } E[\log f_t(Y_t|Y_{t-1},\theta)-\log f_t(Y_t|Y_{t-1},\theta_0)] \leq -\varepsilon_2\}. \quad [4.11]$$

Proof:

$$E[\log f_t(Y_t|Y_{t-1},\theta)-\log f_t(Y_t|Y_{t-1},\theta_0)]$$

$$= p_{t-1,0}\left[\pi_{t0}(1)\ \log\frac{\pi_t(1)}{\pi_{t0}(1)} + (1-\pi_{t0}(1))\ \log\frac{1-\pi_t(1)}{1-\pi_{t0}(1)}\right]$$

$$+ (1-p_{t-1,0})\left[\pi_{t0}(0)\ \log\frac{\pi_t(0)}{\pi_{t0}(0)} + (1-\pi_{t0}(0))\ \log\frac{1-\pi_t(0)}{1-\pi_{t0}(0)}\right]$$

Assumption [A1] confines the components of $\underline{\pi}_{t0}$ to the interval $[p_{min},1-p_{min}]$, so the expected value of the log likelihood ratio is bounded above by

$$\sup_{p_{min}\leq s\leq 1-p_{min}} p_{min}\left[s\ \log\frac{s-\delta}{s} + (1-s)\ \log\frac{1-s+\delta}{1-s}\right].$$

This inequality is true because for fixed s the quantity in brackets is the negative of a Kullback-Leibler distance, so it is an increasing function of $|\delta|$. Because it is a continuous function of s that is negative on a compact set, it achieves some maximum negative value on that set. Applying Lemma 1 concludes the proof.

Theorem. Let $\theta_0 = (\log \varphi_0, \beta_0)$ be the true value of θ and let C be any compact subset of the parameter space that does not contain θ_0. Then assumptions [A1] and [A2] imply

$$\lim_{n\to\infty} \sup_{\theta\in C} \frac{f_2(Y_2|Y_1,\theta) \cdots f_n(Y_n|Y_{n-1},\theta)}{f_2(Y_2|Y_1,\theta_0) \cdots f_n(Y_n|Y_{n-1},\theta_0)} = 0 \quad \text{w.p. 1.}$$

Proof: Note that the above event is equivalent to

$$\lim_{n\to\infty} \sup_{\theta\in C} \sum_{t=2}^{n} \log f_t(Y_t|Y_{t-1},\theta) - \log f(Y_t|Y_{t-1},\theta_0) = -\infty. \qquad [4.12]$$

I will show that this equality holds almost surely.

Pick any θ in C. Define

$$f_t(Y_t|Y_{t-1},\theta,\rho) = \sup_{|\theta'-\theta|<\rho} f_t(Y_t|Y_{t-1},\theta')$$

and

$$U_t(\theta,\rho) = \log f_t(Y_t|Y_{t-1},\theta,\rho) - \log f_t(Y_t|Y_{t-1},\theta_0).$$

$U_t(\theta,\rho)$ is an upper bound on the log of the likelihood ratio in a neighborhood of θ. More specifically, for all $\theta' \varepsilon \{\theta': |\theta-\theta'|<\rho\}$, $U_t(\theta,\rho) > \log f_t(Y_t|Y_{t-1},\theta') - \log f_t(Y_t|Y_{t-1},\theta_0)$.

There are two cases: either $\varphi \neq \varphi_0$, or $\varphi = \varphi_0$ but $\beta \neq \beta_0$. In the second case lemmas 1 and 2 hold. If $\varphi \neq \varphi_0$, then there is some positive δ larger than the minimum distance between the curve

$$\{\underline{\pi} = (\pi(0),\pi(1)): \varphi = \pi(1)(1-\pi(0))/\pi(0)(1-\pi(1)),\ p_{min} \leq \pi(1) \leq 1-p_{min},$$
$$p_{min} \leq \pi(0) \leq 1-p_{min}\}$$

and the curve

$$\{\underline{\pi}=(\pi(0),\pi(1)):\ \varphi_0=\pi(1)(1-\pi(0))/\pi(0)(1-\pi(1)),\ p_{min}\leq\pi(1)\leq 1-p_{min},$$

$$p_{min}\leq\pi(0)\leq 1-p_{min}\}.$$

This is most easily seen by examining Figure 1.1, and is proved by noting that these curves are disjoint compact sets. Then for all t, $d(\underline{\pi}_t,\underline{\pi}_{t0})\geq\varepsilon_1$ for some ε_1. But this implies the conclusion of Lemma 1, so Lemma 2 holds for this case as well.

Pick any positive ε small enough so that lemma 2 applies, and let η and T be as in that lemma. Let I be the set of time indices such that

$$E[\log f_t(Y_t|Y_{t-1},\theta) - \log f_t(Y_t|Y_{t-1},\theta_0)] < -\varepsilon_2.$$

By continuity of the function $E[\log f_t(Y_t|Y_{t-1},\theta')]$ on the compact set $\{(\theta',X_{t-1},X_t):\ \theta'\in C,\ |X_{t-1}|\leq M,\ |X_t|\leq M\}$, and therefore by uniform continuity on that set, there exists a positive $\rho_1=\rho_1(\theta_0,\theta)$ such that for all $t\in I$, $E[U_t(\theta,\rho)]\leq -\varepsilon_2/2$ if $\rho<\rho_1$.

Since $E[\log f_t(Y_t|Y_{t-1},\theta) - \log f_t(Y_t|Y_{t-1},\theta_0)]\leq 0$ for all t, if follows, again by uniform continuity, that there exists a positive $\rho_2=\rho_2(\theta_0,\theta)$ such that for all $t\notin I$, $E[U_t(\theta,\rho)] \leq \eta\varepsilon_2/4(1-\eta)$ if $\rho<\rho_2$. Hence for $n\geq T$ and $\rho_\theta\leq\min(\rho_1(\theta_0,\theta),\rho_2(\theta_0,\theta))$,

$$\sum_{t=2}^{n} E[U_t(\theta,\rho)] \leq -\frac{\varepsilon_2}{2}n\eta + \frac{\eta\varepsilon_2}{4(1-\eta)}n(1-\eta) = -n\eta\varepsilon_2/4\ ,$$

which approaches $-\infty$ as $n\to\infty$.

By compactness, there is a finite covering of C by sets $\{\theta:\ |\theta-\theta_j|<\rho_{\theta j}\}$, $j=1,\ldots,k$. For every $\theta\in C$ there is a $j\leq k$ such that

$$\sum U_t(\theta_j,\rho_{\theta j}) > \sum \log f_t(Y_t|Y_{t-1},\theta) - \log f_t(Y_t|Y_{t-1},\theta_0),$$

so $\{U_t(\theta_j,\rho_{\theta j})\}$, j=1,...,k, form a finite collection of random sequences that bound all the log likelihood ratios for $\theta \varepsilon C$. Therefore to prove the theorem it is sufficient to prove that for every j, $\sum U_t(\theta_j,\rho_{\theta j}) \to -\infty$ almost surely.

Because $\{Y_t\}$ is a *-mixing process, $\{U_t\}$ is also a *-mixing process. Because it is a continuous function on a compact set, it is bounded, so

$$\sum_{n=1}^{\infty} n^{-2} E[(U_n - EU_n)^2] < \infty \quad \text{and} \quad \sup_n n^{-1} \sum_{t=1}^{n} E |U_t - EU_t| < \infty .$$

Therefore by Theorem 2.20 of Hall and Heyde (1980), $n^{-1} \sum [U_t - EU_t] \to 0$ almost surely. But $\sum EU_t \to -\infty$, so the theorem is proved.

Assumption 2 above requires that a non-negligible proportion of (X_{t-1},X_t) be at least some minimal distance from each of a certain family of manifolds. Unfortunately a smooth manifold could be put through any finite collection of points, so this condition is hard to check. A closer examination of the manifolds may be useful.

Let $\rho=\rho_0$, the true value of the odds ratio. The assumption requires that if β_0 is the true value of the coefficient vector, for any other vector β a non-negligible proportion of the time (X_{t-1},X_t) lie at least some minimal distance from the manifold on which $\underline{\pi}_t=\underline{\pi}_{t0}$. Given $X_{t-1}'\beta_0$ and $X_t'\beta_0$, the marginal probabilities $p_{t-1,0}$ and $p_{t,0}$ are determined, and therefore $\underline{\pi}_{t0}$ is determined. But reference to Figure 1.1 clearly indicates that for any value of p_{t-1} there is only one value of p_t that produces a given $\underline{\pi}_t$. Therefore given $X_{t-1}'\beta$, $X_t'\beta$ is determined.

The same statments apply to the other linear combinations, so any three of the quantities $(X_{t-1}'\beta_0, X_t'\beta_0, X_{t-1}'\beta, X_t'\beta)$ determine the other. Therefore each of these quantities is a single-valued function of the other three. There are two conditions under which points cannot be restricted to be on this manifold:

1. (X_{t-1},X_t) is a random variable with a non-degenerate distribution, so it is not concentrated along a lower dimensional manifold with probability one.
2. There are integers t and s such that $X_t=X_s$ but either $X_{t-1}\neq X_{s-1}$ or $X_{t+1}\neq X_{s+1}$.

The second condition is likely to hold if the X's can take only a finite collection of values, or if they are the result of an experimental design. The first condition is likely to hold under a variety of circumstances.

The assumption is somewhat stronger; it requires that a non-negligible proportion of the (X_{t-1},X_t) be bounded away from the manifold. This condition would likely be satisfied if the X's take finitely many values or come from an experimental design. It would also hold if the $\{X_t\}$ were independent and identically distributed with a non-degenerate distribution, and would probably hold under milder conditions as long as the dependence is not too great and the distributions do not converge to a degenerate distribution.

4.3 Comparison of Estimators

Using the expression for the first derivatives and the information matrix from section 1 of this chapter, it is not difficult to find the maximum likelihood estimates of φ and β by Fisher's scoring method. However the results in the previous chapter suggest the ordinary logistic estimates of β may be adequate. In this section I perform a simulation to compare estimators of β and φ.

In this simulation I generated samples of 100 observations Y_t for which the marginal probabilities p_t satisfied $\log (p_t/(1-p_t)) = 1 + x_t$, where the $\{x_t\}$ were independent standard normal random variables. I generated 200 such samples for each of 21 values of φ equally spaced on the log scale between 0.1 and 10.

For each sample I estimated the coefficients both by ordinary logistic regression and by maximum likelihood estimation for the serial dependence model. I estimated the log odds ratio by three methods:

1. Unrestricted maximum likelihood estimation for the serial dependence model (UMLE).
2. One iteration of the scoring method, starting with the ordinary logistic coefficient estimates and with φ=1 (1STEP). (This is equivalent to setting the score statistic, defined in Chapter 5, equal to its expected value and solving for φ.)
3. Restricted maximum likelihood estimation, with the coefficients constrained to their ordinary logistic estimates (RMLE).

The simulation results are summarized in Tables 4.1-2 and Figures 4.1-3.

Table 4.1 shows the correlation between those estimators that estimate the same quantity. In each case the correlation seems to vary with the odds ratio, with lower correlations for odds ratios far from 1. Most correlations exceed 0.9 when the odds ratio is not far from 1. The correlation between RMLE and UMLE is quite near 1 for a wide range of odds ratios, so this is an indication that if the odds ratio is not expected to be extreme a priori, the restricted maximum likelihood estimator, which is simpler to compute, may be as good as the unrestricted maximum likelihood estimator.

If the odds ratio is known a priori to be neither zero nor infinity, the 1STEP estimator has an additional advantage: it always gives a finite estimate of the odds ratio. Perfect association occurred in many samples, so the RMLE and UMLE estimates of the log odds ratio are infinite.

Table 4.2 shows the observed bias and standard deviation for the coefficient estimates. As was shown in the previous chapter, the standard deviations of the intercept increase with the odds ratio, and this phenomenon is apparent here as well. The bias of the slope estimates is smaller for most of the maximum likelihood estimates, but there is no noticeable pattern to the other quantities. Because the biases are small in comparison with the standard deviations, a much larger sample would be required to test for a significant difference between the logistic and maximum likelihood biases.

Table 4.1. Correlations of Estimates

For each value of the odds ratio, the table gives the correlation between the given pair of estimates in a sample of size 200. The last column gives the number of observations with perfect association; for these samples the UMLE and RMLE values are zero or infinity. These observations were excluded in computing the correlations.

Odds Ratio	Constant	Slope	1STEP/ RMLE	1STEP/ UMLE	RMLE/ UMLE	Perfect Association
.10	.986	.869	.865	.693	.733	10
.13	.989	.903	.849	.723	.925	17
.16	.988	.877	.918	.848	.944	0
.20	.990	.894	.894	.779	.943	4
.25	.993	.969	.954	.944	.981	1
.32	.991	.952	.952	.898	.970	0
.40	.993	.959	.964	.791	.868	3
.50	.996	.962	.985	.981	.998	0
.63	.990	.981	.981	.981	.999	0
.79	.999	.981	.995	.995	.999	0
1.00	.997	.962	.995	.994	.999	0
1.26	.999	.991	.996	.994	.998	0
1.58	.993	.989	.994	.993	1.000	0
2.00	.997	.962	.990	.955	.969	0
2.51	.993	.989	.982	.981	.999	0
3.16	.995	.961	.986	.959	.972	0
3.98	.990	.965	.972	.935	.971	0
5.01	.989	.923	.965	.859	.908	0
6.31	.985	.925	.939	.936	.998	0
7.94	.979	.927	.937	.779	.853	1
10.00	.987	.897	.897	.717	.818	5

Table 4.2. Observed Bias and Standard Deviation

The elements in the table are the sample bias and standard deviation of the coefficient estimates in the simulation. Observations with perfect association are omitted.

	Constant				Slope			
Odds	Logistic		MLE		Logistic		MLE	
Ratio	Bias	Std dev	Bias	Std dev	Bias	Std dev	Bias	Std dev
.10	.0047	.208	.0024	.201	.0211	.309	.0153	.273
.13	-.0026	.180	-.0064	.176	.0694	.327	.0550	.289
.16	.0224	.212	.0176	.207	.0057	.288	.0125	.275
.20	-.0088	.200	-.0121	.198	.0352	.254	.0249	.265
.25	.0409	.247	.0414	.253	.0793	.308	.0680	.320
.32	.0476	.235	.0492	.242	.0740	.299	.0638	.316
.40	.0185	.211	.0169	.212	.0378	.355	.0298	.339
.50	.0314	.227	.0299	.224	.0471	.291	.0394	.296
.63	.0290	.232	.0295	.236	.0345	.347	.0363	.342
.79	.0369	.274	.0371	.276	.0360	.286	.0381	.290
1.00	.0451	.285	.0471	.288	.0605	.294	.0681	.307
1.26	.0323	.281	.0297	.281	.0523	.323	.0504	.330
1.58	.0312	.306	.0289	.306	.0077	.300	.0073	.294
2.00	.0827	.310	.0842	.311	.0676	.310	.0676	.319
2.51	.0547	.299	.0525	.300	.0234	.276	.0218	.276
3.16	.0213	.310	.0155	.308	.0452	.302	.0355	.307
3.98	.0687	.317	.0657	.319	.0911	.307	.0877	.297
5.01	.0649	.378	.0574	.373	.0474	.317	.0330	.285
6.31	.0672	.426	.0590	.431	.0617	.330	.0675	.308
7.94	.0150	.327	.0139	.322	.0408	.237	.0388	.230
10.00	.0416	.395	.0351	.383	.0740	.283	.0546	.265

Figure 4.1. Stem and Leaf Display of Log Odds Ratio Estimates
True Log Odds Ratio is 0.0

```
One-step Estimate.  Values off plot: -1.8 -1.6 -1.3 1.3
                                                  1.4
  5   -1.22
  8   -1.100
 11   -0.988
 24   -0.7777766666666
 41   -0.55555555444444444
 69   -0.333333333333333222222222222
(35) -0.11111111111111111111100000000000000
 96   +0.00000000000000000001111111111111
 65   +0.22222222222222333333333
 41   +0.444444445555555
 26   +0.6666667777777
 13   +0.88889999
  5    1.001

Restricted MLE.  Values off plot: -2.1 -1.6 -1.5 -1.5
                                        -1.5 1.6
  8   -1.111
 14   -0.999888
 25   -0.77777766666
 42   -0.55555555554444444
 71   -0.333333333333333322222222222222
(33) -0.11111111111111111100000000000000000
 96   +0.00000000000000000011111111111111
 64   +0.22222222222223333333333
 41   +0.4444444455555555
 25   +0.666666777777
 13   +0.888888999
  4    1.1
  3    1.22

Unrestricted MLE.  Values off plot: -2.2 -1.8 -1.6 -1.6
                                          -1.5 1.4 1.8
  8   -1.111
 14   -0.999988
 27   -0.7777777666666
 44   -0.55555555554444444
 71   -0.333333333333333222222222222
(33) -0.11111111111111111100000000000000000
 96   +0.00000000000000000011111111111111
 64   +0.22222222222223333333333
 41   +0.444444445555555
 26   +0.6666667777777
 13   +0.888889999
  4    1.1
  3    1.2
```

Figure 4.2. Stem and Leaf Display of Log Odds Ratio Estimators
True Log Odds Ratio is 0.69

```
 One-step Estimate.  Values off plot: -.9 2.2
  5   -0.7766
  6   -0.4
 16   -0.3333333222
 28   -0.111111110000
 46   +0.000000001111111111
 67   +0.222222222333333333
 89   +0.444444444444455555555
(38)  +0.66666666666666666677777777777777777777
 73   +0.888888888888999999999999
 49    1.000000000111111111
 30    1.2222222233
 20    1.44444455555
  9    1.66677
  4    1.99
  2    2.1

Restricted MLE.  Values off plot: -1.0 2.4
  2   -0.8
  5   -0.766
  9   -0.4444
 17   -0.33332222
 28   -0.11111110000
 45   +0.00000000111111111
 71   +0.22222222222333333333333333
 88   +0.44444444555555555
(42)  +0.66666666666666666666666677777777777777777777
 70   +0.888888888899999999999999
 48    1.000000001111111111
 29    1.2222223333
 19    1.445555
 13    1.666677777
  4    1.888

 Unrestricted MLE.  Values off plot: -1.0 2.3 4.7
  2   -0.8
  5   -0.766
  9   -0.4444
 17   -0.33333222
 28   -0.11111110000
 45   +0.00000000111111111
 69   +0.22222222223333333333333
 87   +0.444444444455555555
(38)  +0.66666666666666666677777777777777777777
 75   +0.888888888888869999999999
 51    1.000000001111111111
 34    1.222222223333
 22    1.445
 19    1.666667777
 10    1.889999
  4    2.00
```

Figure 4.3. Stem and Leaf Display of Log Odds Ratio Estimators
True Log Odds Ratio is 2.30

```
One-Step Estimate.  Values off plot: 0.1 0.2 3.6
 4   +0.66
 6   +0.89
11    1.00011
27    1.2222333333333333
43    1.4444444455555555
62    1.6666666666677777777
89    1.88888888888888889999999999
(39)  2.000000000000000000001111111111111111111
67    2.22222222333333333
50    2.44444444555555555
33    2.66666677777777
19    2.888888888999999
 4    3.011

Restricted MLE.  Values off plot: 0.1 0.1 4.3
 3   +0.5
 4   +0.6
 6   +0.89
13    1.0011111
22    1.223333333
35    1.4444455555555
54    1.66666666666777777777
72    1.888888888888999999
93    2.0000000001111111111111
(26)  2.22222222222333333333333333
76    2.44444444444444555555555
54    2.6666666677777777
38    2.8889999999
28    3.000111
22    3.222233333
13    3.444455
 7    3.77
 5    3.8899

Unrestricted MLE.  Values off plot: 5.5 5.5 5.6 5.6 5.8
                                    5.9 5.9 6.7 6.8 7.9 8.8
13.2
 2    +0.11
 6    +0.5689
24     1.000112223333333444
57     1.555556666777777788888888889999999
93     2.000000001112222233333333333344444444
(42)   2.55555555555566666666667777777788889999999
60     3.0000111122233444
44     3.5556777789
34     4.11112233
26     4.555789
20     5.00012222
```

Figures 4.1-3 show stem and leaf plots of the log odds ratio estimators for three values of the odds ratio: 1, 2, and 10. In Figure 4.1, with ψ=1, there seems to be little difference between the estimators based on this display, except the number of extreme values ("values off plot" in the figures) is smallest for 1STEP and largest for UMLE. The medians of the estimators are about the same, and all are close to the true value.

In Figure 4.2 the situation is little different. There is a very large value of UMLE, but otherwise the shapes of the distributions appear to be very similar. Again the medians are about the same, and all are close to the true value.

Figure 4.3 shows a more dramatic difference. The estimator 1STEP retains the bell shape it assumed in the other figures, but RMLE has a noticeably heavier upper tail. The shape of UMLE is even more skewed, with eleven extreme values off the high end of the plot.

The median of 1STEP seems to be a little smaller than the true value, while UMLE is a little larger. However for five observations (not included in the plot) perfect association occurred, so RMLE and UMLE both were infinite.

In summary, this relatively small simulation does not show any advantage to using maximum likelihood estimation in the serial dependence model rather than ordinary logistic coefficient estimates. Using these estimates and ψ=1 as starting values, performing a single iteration of the

scoring method gives a good estimate of the log odds ratio. When ψ is near 1 this estimator is not noticeably different from the others, but when ψ is extreme it avoids the problem of perfect correlation.

If the standard errors of the estimates are also of interest, as is usually the case, then the maximum likelihood estimate remains important. The standard errors produced by this model differ from those produced by the ordinary logistic model, and are closer to reality.

Chapter 5

A Test for Independence

5.1 Introduction

Compared with ordinary logistic regression, maximum likelihood estimation of this model takes a relatively large amount of computer time and requires specialized software. It would be useful therefore to be able to test for dependence without having to do the maximum likelihood computations. In some cases it may not be thought necessary to compute the maximum likelihood estimates if a preliminary test does not reject the hypothesis of independence.

Such a test can be based on the score statistic, which requires maximization only over the subset of the parameter space that satisfies the null hypothesis. Since independence implies $\rho=1$, testing for independence requires computing the ordinary logistic estimates, as they maximize the likelihood subject to this restriction. As a result the score test can be performed with little more computational effort than that required by logistic regression.

In ordinary linear regression, a test for serial correlation can be based on the sample autocorrelation function of the residuals, or equivalently on the Durbin-Watson statistic. A test based on the latter was developed in a series of papers by Durbin and Watson (1950, 1951, 1971). As I will

show below, the score test for this model is based on the sample autocovariance of the data.

The score test is described by Rao (1973) and by Cox and Hinkley (1974). Let U be the score function and i the information matrix. To test a portion ψ of a parameter vector (ψ,β), Cox and Hinkley define the score statistic W by

$$W = U_{\psi}' \, i^{\psi\psi} \, U_{\psi} , \qquad [5.1]$$

where U_{ψ} is the portion of the score function corresponding to ψ and $i^{\psi\psi}$ is the corresponding submatrix of the inverse of the information matrix. When ψ is a scalar it is more convenient to retain the sign by using instead

$$W^{1/2} = U_{\psi}(i^{\psi\psi})^{1/2}. \qquad [5.2]$$

I derive the score statistic for this problem in section 2. In section 3 I give the asymptotic distribution of W under the null hypothesis and examine the empirical distribution in finite samples. In later sections I consider the distribution under alternative hypotheses.

5.2 Derivation

The score function is defined as the derivative of the log likelihood. Expressions for the derivatives appear on pages 41-42. Since this is a test for independence, these expressions are to be evaluated at $\psi=1$. Therefore $\alpha_t = p_t p_{t-1}$ and the score functions become

$$U_{\psi} = \frac{\partial L}{\partial \psi} = \sum_{t=2}^{n} (Y_t - p_t)\,(Y_{t-1} - p_{t-1}) \qquad [5.3]$$

$$U_\beta = \frac{\partial L}{\partial \beta} = \sum_{t=1}^{n} (Y_t - p_t)\, x_t \,. \qquad [5.4]$$

Note that since the unconditional expectation of Y_t is p_t, the score function for the odds ratio has the form of the autocovariance of the $\{Y_t\}$.

The statistic also involves the information matrix. The three components evaluated at $\varphi=1$ are

$$-E\left[\frac{\partial^2 L}{\partial \varphi^2}\right] = \sum p_t(1-p_t)p_{t-1}(1-p_{t-1}) \,. \qquad [5.5]$$

$$-E\left[\frac{\partial^2 L}{\partial \beta^2}\right] = \sum p_t(1-p_t)x_t x_t' \qquad [5.6]$$

$$-E\left[\frac{\partial^2 L}{\partial \beta\, \partial \varphi}\right] = 0. \qquad [5.7]$$

Therefore the complete information matrix can be written

$$i = \begin{bmatrix} i_{\varphi\varphi} & i_{\varphi\beta} \\ i_{\beta\varphi} & i_{\beta\beta} \end{bmatrix}$$

$$= \sum_{t=1}^{n} p_t(1-p_t) \begin{bmatrix} p_{t-1}(1-p_{t-1}) & 0 \\ 0 & x_t x_t' \end{bmatrix} , \qquad [5.8]$$

if p_0 is taken to be equal to 0.

The score statistic is therefore

$$W = \frac{[\sum (Y_t - \tilde{p}_t)(Y_{t-1} - \tilde{p}_{t-1})]^2}{\sum \tilde{p}_t(1-\tilde{p}_t)\, \tilde{p}_{t-1}(1-\tilde{p}_{t-1})} \,, \qquad [5.9]$$

where $\tilde{p}_t = 1/(1+\exp(-x_t^T\tilde{\beta}))$ and $\tilde{\beta}$ is the estimate of β in the ordinary logistic model. Since this is a test on a scalar parameter, the other statistic can also be defined:

$$W^{1/2} = \frac{[\Sigma\ (Y_t-\tilde{p}_t)(Y_{t-1}-\tilde{p}_{t-1})]}{[\Sigma\ \tilde{p}_t(1-\tilde{p}_t)\ \tilde{p}_{t-1}(1-\tilde{p}_{t-1})]^{1/2}} \qquad [5.10]$$

5.3 Distribution of the score statistic under the null hypothesis

According to asymptotic theory, the distribution of W under the null hypothesis of independence should be central chi-square with one degree of freedom, and the asymptotic distribution of $W^{1/2}$ should be standard normal. To examine the distribution of W for finite samples I used a simulation. Figure 5.1 shows a normal probability plot of a sample of $W^{1/2}$ generated by the following procedure:

(1) generate $X_1,\ldots,X_{100}$ independent standard normal random variates

(2) generate $Y_1,\ldots,Y_{100}$ independent Bernoulli random variates with probability p_t of success satisfying $\log(p_t/(1-p_t)) = 1+X_t$

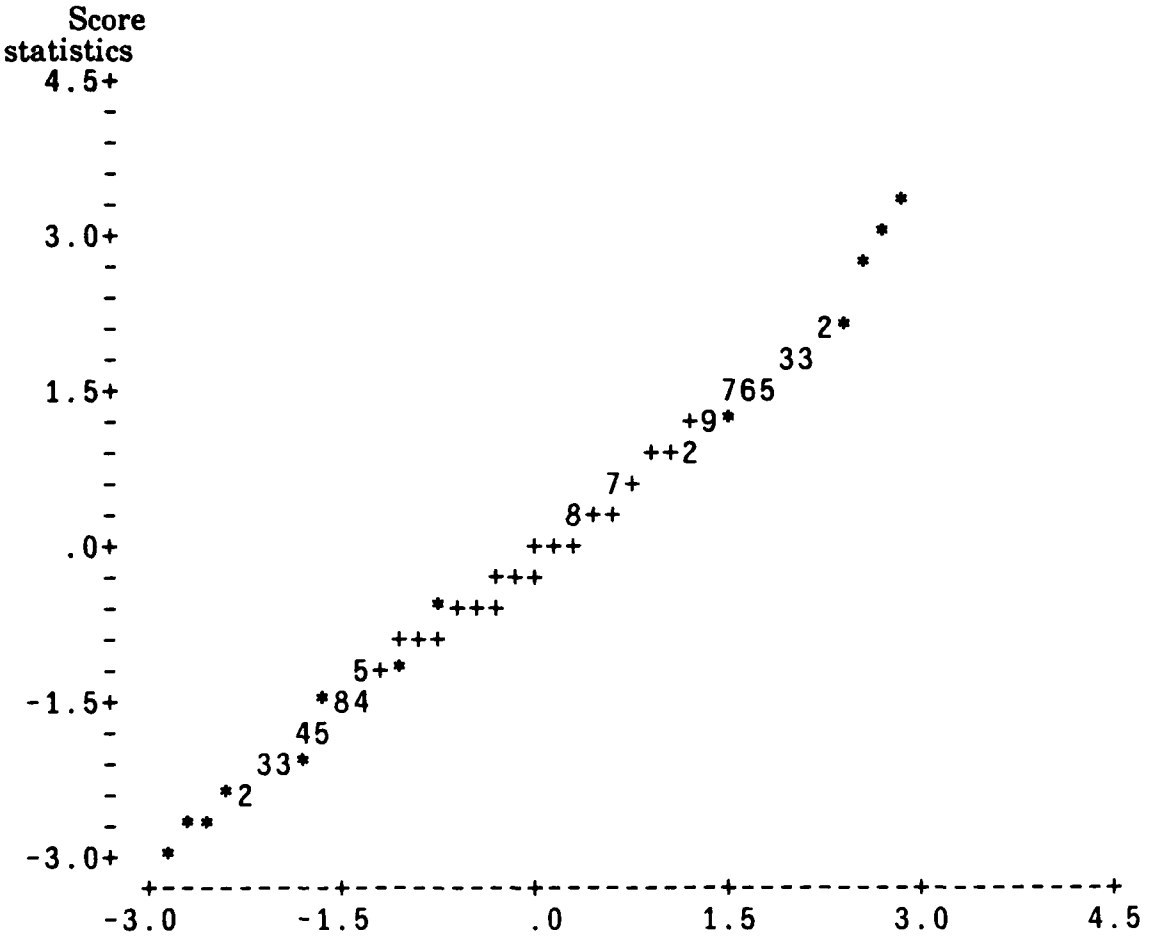
(3) perform ordinary logistic regression and compute $W^{1/2}$

I generated a sample of 400 score statistics by this procedure, using the same $\{X_t\}$ each time. From the plot, the sample seems consistent with a standard normal distribution.

Since W is a test statistic, its upper tail behavior is of interest. For example, (0.10)(400) = 40 of the score statistics could be expected to fall above the 0.90 point of the chi-square distribution, or 2.706. In this sample 48 were observed above the critical value. If the distribu-

Figure 5.1. Normal Probability Plot for Score Statistics

Generated with $\psi = 1$



tion is correct, the number of significant statistics should have a binomial distribution with mean 40 and standard deviation $[(0.1)(0.9)(400)]^{1/2}$ = 6, so a difference of 8 from the expected value is not significant at the 0.05 level.

The Kolmogorov-Smirnov distance between the empirical cumulative distribution of $W^{1/2}$ and the standard normal cumulative distribution function is 0.049, a value that is below the 0.95 point of the distribution of the Kolmogorov-Smirnov statistic, or $1.36n^{-1/2}$ = 0.068.

5.4 Distribution for nearby alternatives: first order approximation

To get some idea of the power of the test, it would be useful to find the distribution of the score statistic when the odds ratio is not equal to one. It is simpler to work directly with the odds ratio, but when considering alternative hypotheses it seems more natural to use $\lambda=\log \psi$ as the parameter measuring dependence. Unlike ψ, λ can vary in either direction without limit. As will be seen below, the effect of alternative values of λ on the distribution of the score statistic depends on the magnitude of λ but is independent of its sign, or nearly so.

The asymptotic distribution of a score statistic under alternative hypotheses is given, for example, by Cox and Hinkley (1974). Suppose the score statistic W is computed for a sample of size n. With the null hypothesis H_0: $\lambda=0$, if a sequence of alternatives H_n: $\lambda=\delta n^{-1/2}$ is to be considered, then the asymptotic distribution of W is non-central chi-square with one degree of freedom and with non-centrality parameter $\delta^2 i_{\lambda\lambda}/n$. (This is

true here because the information matrix is block diagonal. In general $i_{\lambda\lambda}$ should be replaced by $(i^{\lambda\lambda})^{-1}$, where $i^{\lambda\lambda}$ is the corresponding submatrix of the inverse of i.) Equivalently $W^{1/2}$ has an asymptotic normal distribution with mean $\delta(i_{\lambda\lambda}/n)^{1/2}$ and variance 1.

As a basis for comparison, I generated twenty additional samples of 400 score statistics as above for various values of the odds ratio. I used the same set of $\{X_t\}$ in each case. I calculated the observed power function in each case as the proportion of W values larger than 2.706, the 0.90 point of the central chi-square distribution with one degree of freedom. These values are those labeled "observed power" in Figure 5.2. For each value θ of the observed power, I calculated a 95% confidence interval for the true power as $\theta \pm 1.96(\theta(1-\theta)/400)^{1/2}$, and these confidence intervals also appear in Figure 5.2.

The power given by asymptotic theory is simply P[V>2.706], where V has a chi-square distribution with one degree of freedom and with non-centrality parameter $\lambda^2 i_{\lambda\lambda} = 3.1(\log \psi)^2$. In Figure 5.3, this curve is superimposed on the simulation results.

Clearly there is some lack of fit, since eight of the twenty-one confidence intervals do not contain the value on the curve. Qualitatively, though, the curve does seem to predict the observed power pretty well. For $0.3 \leq \psi \leq 2.5$, the intervals do contain the curve. It is not surprising to see a lack of fit for more extreme values, since the curve is obtained

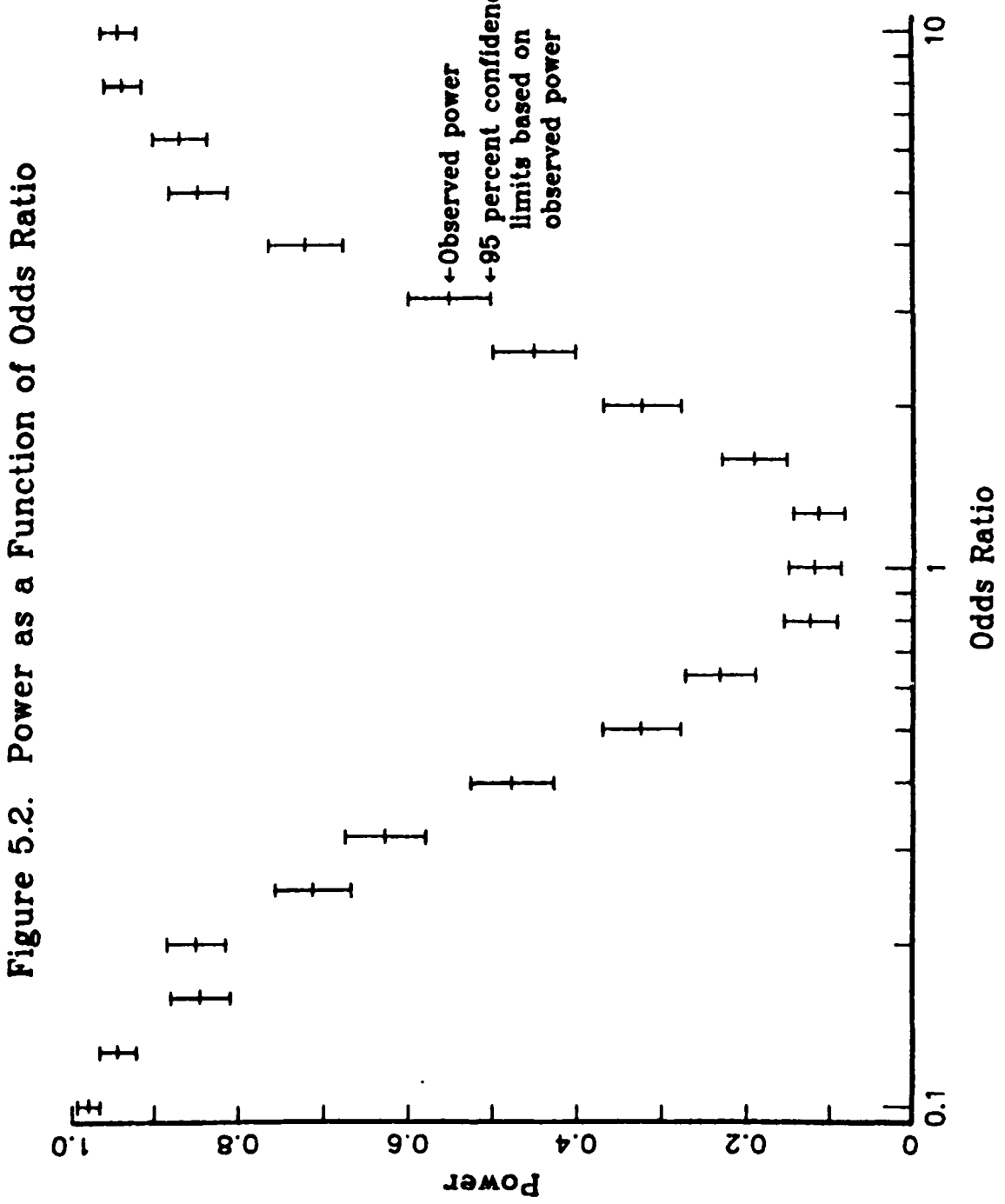


Figure 5.2. Power as a Function of Odds Ratio

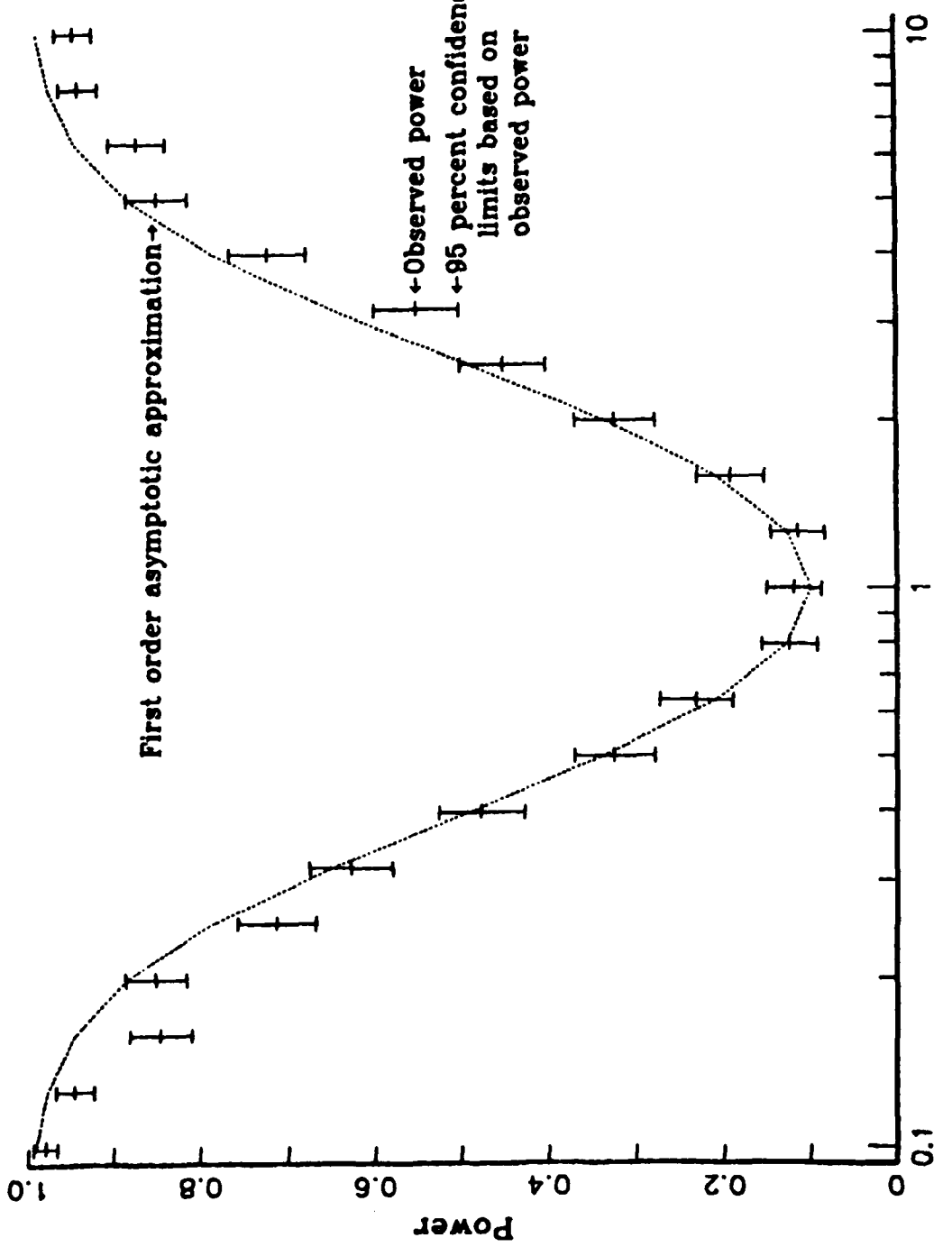


Figure 5.3. Power as a Function of Odds Ratio

by an asymptotic calculation valid for nearby alternatives.

The above procedure sheds some light on the upper tail of the distribution of the score statistic, but not on the bulk of the distribution of W. I used the Kolmogorov-Smirnov statistic as a summary of the difference between the cumulative distribution functions of the empirical and asymptotic distributions, and the results appear in Table 5.1. Column 1 contains the odds ratio, and column 2 contains $n^{1/2}$=20 times the Kolmogorov-Smirnov distance D between the two distributions. (The other columns are explained below.) Values larger than the upper 95% point of the distribution of $n^{1/2}D$, or 1.36, are marked with an asterisk. The results here are similar to those above; there is no significant lack of fit for $.4 \leq \varphi \leq 2.5$.

The following sections describe two attempts to improve the accuracy of the power curve. In the first attempt I obtain a higher order approximation to the distribution of W using the results of Harris and Peers (1980). In the second I use more informal techniques to examine the deviation of the simulated $W^{1/2}$ from its theoretical distribution and I find an empirical adjustment that improves the approximation to the observed power.

5.5 Distribution for nearby alternatives: higher order approximation

One approach toward improving the fit to the observed power function is to find a higher order approximation to the distribution of the score statistic. Peers (1971) gave such an approximation for simple tests, and Harris and Peers (1980) extended the results to composite tests.

Table 5.1. Normalized Kolmogorov-Smirnov Distances Between the Empirical and Fitted Distributions

For each value of the odds ratio the other columns give $n^{1/2}=20$ times D, the Kolmogorov-Smirnov distance between the empirical distribution of the corresponding sample and the fitted distribution. Values larger than the 0.95 point of the null distribution of $n^{1/2}D$, or 1.36, are marked with an asterisk.

Odds Ratio	First Order Approximation	Higher Order Approximation	Empirical Approximation
.10	7.903*	2.456*	1.330
.13	5.842*	1.076	1.017
.16	4.931*	.803	1.402*
.20	2.352*	1.609*	.810
.25	2.342*	1.350	.785
.32	1.713*	.985	.759
.40	.734	1.571*	.931
.50	.860	1.080	.861
.63	1.076	1.331	1.317
.79	1.338	1.304	1.205
1.00	.972	.972	.972
1.26	.596	.637	.593
1.58	1.255	1.341	1.234
2.00	.929	1.461*	.816
2.51	.881	1.794*	.590
3.16	1.712*	3.207*	1.012
3.98	2.314*	4.391*	1.144
5.01	2.888*	6.201*	.769
6.31	3.718*	7.950*	.560
7.94	3.603*	9.032*	1.309
10.00	4.361*	11.154*	2.826*

Because the derivation is lengthy, I present the results first.

The density of the score statistic W to order $n^{-1/2}$ can be written as a linear combination of non-central chi-square densities, each with the same non-centrality parameter but with different degrees of freedom. More specifically the density is

$$g(w) = f(w;1,\rho) + n^{-1/2} \sum_{j=0}^{3} c_j \, f(w;1+2j,\rho) + O(n^{-1}), \qquad [5.11]$$

where $f(.;j,\rho)$ is the density of a chi-square random variable with j degrees of freedom and with non-centrality parameter ρ. I will give the values of ρ and the c_j's below.

Several points can be made about this approximation. First, the first term on the right hand side is the usual (first order) approximation to the distribution of W under alternative hypotheses. Second, under the null hypothesis all the c_j's are zero, so both approximations lead to the same distribution in this case. Third, this is simply an approximation to an asymptotic distribution, and it will not necessarily integrate to one in finite samples.

Figure 5.4 contains the power function for this higher order approximation as well as the power function examined earlier. Any improvement here is marginal. The new curve seems to give a better fit to the observed power for $\varphi<1$, but it gives a poorer fit for $\varphi>1$. Because it is not a true density and does not integrate to one, it gives values of the "power func-

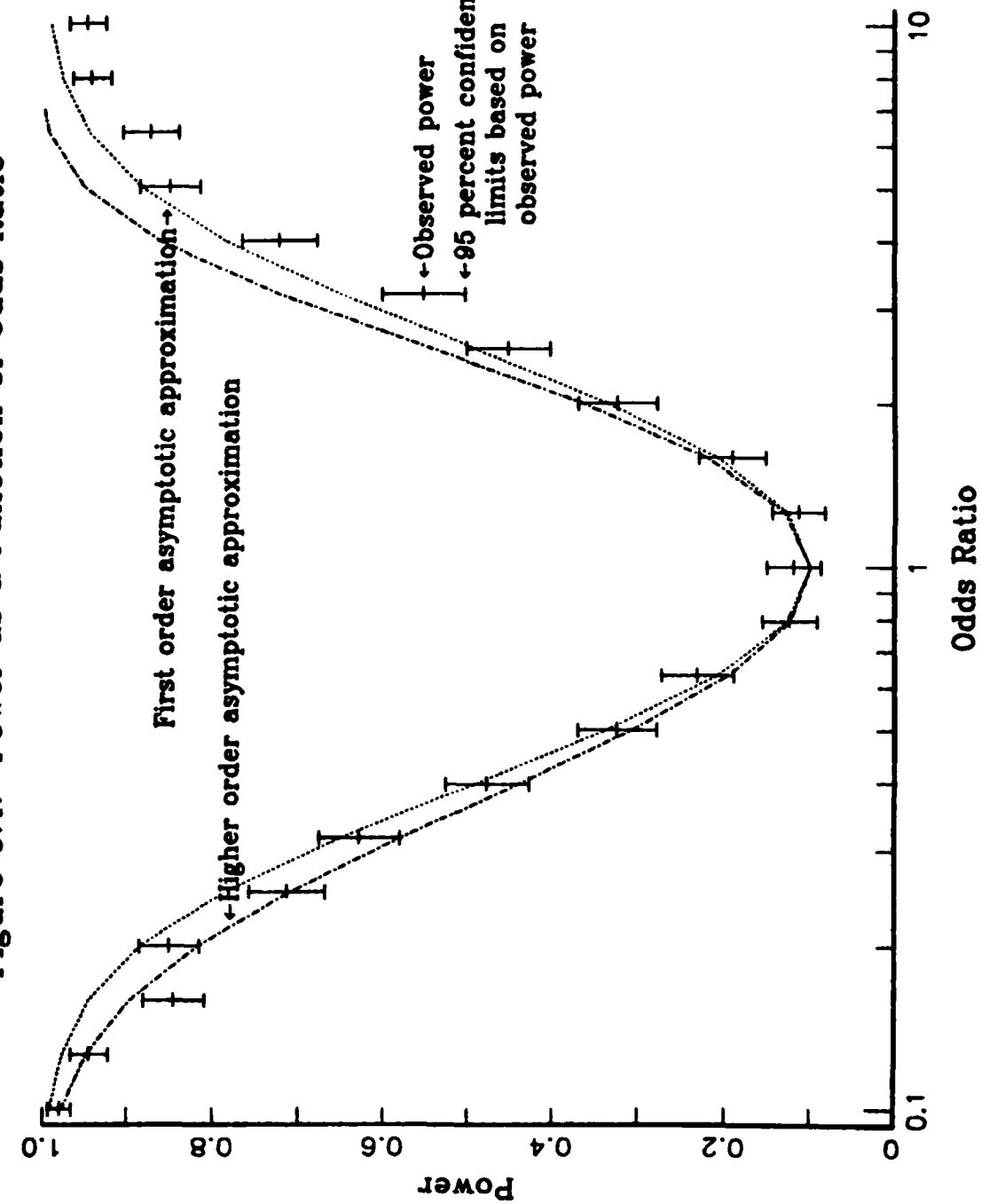


Figure 5.4. Power as a Function of Odds Ratio

tion" that are larger than one for $\varphi \geq 8$. Since the first order approximation has the advantage of simplicity, there seems to be no reason to prefer the more complex approximation.

The third column of Table 5.1 contains the normalized Kolmogorov-Smirnov distances between the empirical distribution of the sample and the higher order asymptotic distribution. It, too, shows mixed results, with an improved fit for $\varphi < 1$ but a poorer fit for $\varphi > 1$.

The remainder of this section consists of calculations of the quantities needed to apply the Harris and Peers results to this model.

Though it is more natural to use $\lambda = \log \varphi$ as the parameter of interest, it is more convenient to work with φ. To convert the results for use with λ requires use of the chain rule to derivatives of up to third order. For any function $u(\varphi)$,

$$\frac{\partial u}{\partial \lambda} = \varphi \frac{\partial u}{\partial \varphi}$$

$$\frac{\partial^2 u}{\partial \lambda^2} = \varphi^2 \frac{\partial^2 u}{\partial \varphi^2} + \varphi \frac{\partial u}{\partial \varphi}$$

$$\frac{\partial^3 u}{\partial \lambda^3} = \varphi^3 \frac{\partial^3 u}{\partial \varphi^3} + 3\varphi^2 \frac{\partial^2 u}{\partial \varphi^2} + \varphi \frac{\partial u}{\partial \varphi} \quad .$$

Following the notation of Harris and Peers, define

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \lambda \\ \beta \end{bmatrix}, \qquad u_i = \frac{1}{n^{1/2}} \frac{\partial L}{\partial \theta_i} \quad ,$$

$$u_{ij} = \frac{1}{n} \frac{\partial^2 L}{\partial\theta_i \partial\theta_j}, \qquad u_{ijk} = \frac{1}{n^{3/2}} \frac{\partial^3 L}{\partial\theta_i \partial\theta_j \partial\theta_k},$$

$$k_{ij} = E[u_{ij}], \quad k_{i,j} = E[u_i u_j], \quad k_{i,j,k} = n^{1/2} E[u_i u_j u_k],$$

$$k_{i,jk} = n^{1/2} E[u_i u_{jk}], \quad k_{ijk} = n^{1/2} E[u_{ijk}],$$

where L is the log likelihood. All expectations are taken with φ=1 and with the true value of β. All the k's are O(1). Relations between these quantities are given by Harris and Peers; in addition to the familiar relationship

$$k_{i,j} + k_{ij} = 0,$$

there is a relationship between the remaining expectations:

$$k_{ijk} + k_{i,jk} + k_{j,ik} + k_{k,ij} + k_{i,j,k} = 0.$$

Let the symbol K represent the matrix of $k_{i,j}$'s; it is simply the information matrix calculated earlier. Let subscripts on K refer to the corresponding submatrix, for example K_{22} is the submatrix of K corresponding to β. Let a dot subscript refer to the entire dimension of K, so $K_{1.} = [K_{11}\ K_{12}]$. Define triply-subscripted K's similarly, so for example $K_{2..}$ is a three dimensional array of $k_{i,jk}$, with i>1.

Let $\varepsilon = n^{1/2} \log \varphi$ measure the distance between the hypothesized and true values of λ. Define the following:

$$\eta = \begin{bmatrix} 1 \\ -K_{22}^{-1} K_{21} \end{bmatrix} \varepsilon , \qquad J = \begin{bmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{bmatrix} ,$$

$$B = K^{-1} - J = \begin{bmatrix} \frac{n}{\sum p_t(1-p_t)p_{t-1}(1-p_{t-1})} & 0 \\ 0 & 0 \end{bmatrix} ,$$

Then the probability density of the score statistic is given by equation [5.11] with non-centrality parameter

$$\rho = \eta^T K \eta = \lambda^2 / \sum p_t(1-p_t)p_{t-1}(1-p_{t-1})$$

and with coefficients

$$c_0 = \frac{1}{6} \left[\varepsilon^3(k_{1,1,1}-k_{111}) + 3\varepsilon^3(k_{111}+k_{1,11}) - 3\varepsilon(K_{..1}+2K_{...1})*J \right]$$

$$c_1 = \frac{1}{6} \left[\varepsilon^3(k_{111}-2k_{1,1,1}) - 3\varepsilon^3(k_{111}+k_{1,11}) - 3\varepsilon k_{1,1,1} n\rho/\lambda^2 \right.$$
$$\left. + 3\varepsilon(K_{..1}+2K_{...1})*J \right]$$

$$c_2 = \frac{1}{2} \varepsilon k_{1,1,1} n\rho/\lambda^2$$

$$c_3 = \frac{1}{6} \varepsilon^3 k_{1,1,1} .$$

The notation A*B used in the expressions for c_0 and c_1 means $\sum_{ij} A_{ij} B_{ij}$.

I present expressions for the remaining quantities without proof:

$$K_{1,1,1} = n^{-1} \sum p_{t-1}(1-p_{t-1})(1-2p_{t-1})\ p_t(1-p_t)(1-2p_t)$$

$$K_{111} = -3n^{-1} \sum p_{t-1}(1-p_{t-1})(1-2p_{t-1})\ p_t(1-p_t)(1-2p_t)$$

$$K_{1,11} = K_{1,22} = K_{1,2,2} = K_{2,1,2} = K_{2,2,1} = 0$$

$$K_{2,12} = K_{2,21} = -n^{-1} \sum p_t(1-p_t)\, p_{t-1}(1-p_{t-1})\, x_t x'_{t-1}$$

$$K_{221} = n^{-1} \sum p_t(1-p_t)\, p_{t-1}(1-p_{t-1})\, (x_{t-1}x'_t + x_t x'_{t-1})$$

5.6 Distribution for nearby alternatives: empirical approximation

Lack of fit of $W^{1/2}$ to its asymptotic distribution could take various forms; it could have a normal distribution but with a different mean, a normal distribution with a different mean and variance, or a distribution that is not normal. In this section I will find the nature of the lack of fit. I will apply exploratory techniques rather than large-sample theory to model the lack of fit and improve the power curve.

First suppose the distribution is normal, but the parameter values are not as predicted by asymptotic theory. The sample means and standard deviations of the score statistics $W^{1/2}$ obtained by simulation appear in Table 5.2 and in Figures 5.5 and 5.6, as a function of the odds ratio. The dashed lines in the Figures give the values predicted by theory.

The sample means are quite near the lines for odds ratios near 1, but they are smaller in absolute value for more extreme values of the odds ratio. The sample standard deviations are near 1 when the odds ratio is larger than 1, but decrease as the odds ratio approaches 0. I will attempt to fit these points empirically and see if the fitted parameter values produce a power curve that is closer to the observed power.

Table 5.2. Sample Means and Standard Deviations for the Score Statistics $W^{1/2}$ from the Simulation

For each value of the odds ratio, the table contains the mean and standard deviation for the corresponding sample of 400 simulated score statistics.

Odds Ratio	Mean	Standard Deviation
.10	-3.1401	.811
.13	-2.9747	.854
.16	-2.7089	.857
.20	-2.5578	.891
.25	-2.2370	.872
.32	-1.8835	.891
.40	-1.5818	.950
.50	-1.2058	.927
.63	-.8701	1.03
.79	-.4426	.877
1.00	-.1030	.971
1.26	.3228	1.03
1.58	.7173	1.03
2.00	1.1245	.979
2.51	1.5460	1.03
3.16	1.8816	1.01
3.98	2.1756	1.03
5.01	2.4961	1.00
6.31	2.8428	1.01
7.94	3.1542	1.03
10.00	3.5221	.962

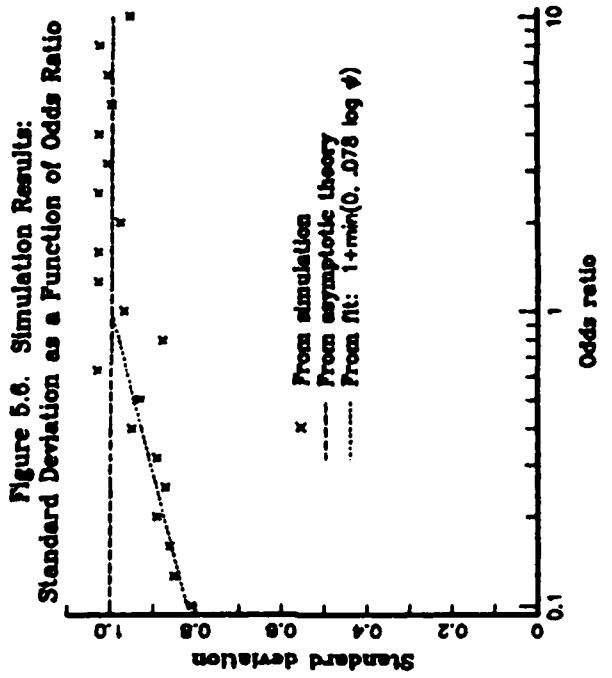


Figure 5.6. Simulation Results:
Standard Deviation as a Function of Odds Ratio

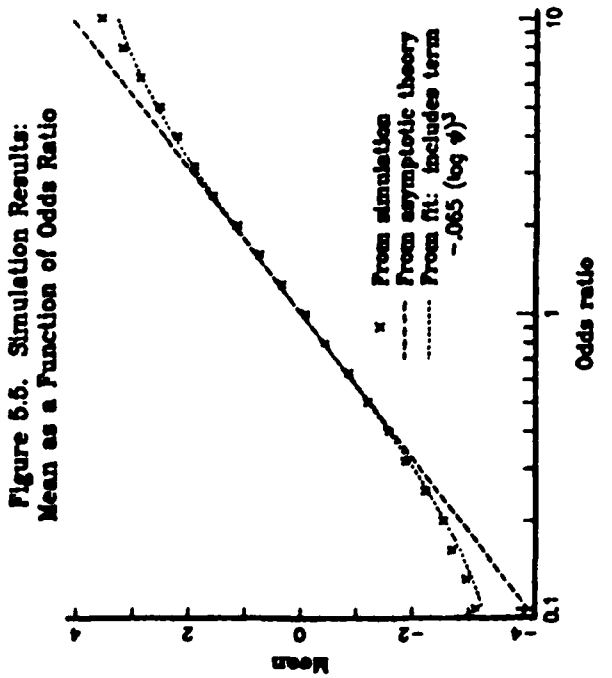


Figure 5.5. Simulation Results:
Mean as a Function of Odds Ratio

Because the sample means appear to be an odd function of the log odds ratio and because the simple linear function fits well for odds ratios near 1, this suggests adding a cubic term in log ψ. A least squares fit with the linear term constrained to $3.1^{1/2}$ (the square root of the information) leads to -0.065 as the coefficient of the cubic term. This is the dotted line in Figure 5.5, and it appears to give a good fit to the sample means.

The sample standard deviations, however, appear to lie on two lines: one at the constant value 1.0 for $\psi>1$ and the other with a positive slope. It seems reasonable to assume continuity, so I fit the second line by least squares subject to the constraint that it pass through the point (1,1). The estimated slope is 0.078. This is the dotted line in Figure 5.6.

Figure 5.7 shows the previous power curves along with one obtained by using parameters given by this empirical fit. The new curve is a marked improvement; it misses only 2 of the 21 confidence intervals.

The last column in Table 5.1 also shows a good fit. Only the score statistics calculated with the odds ratios equal to .16 and 10 produce a normal distribution that is significantly different from the empirical distribution at the 0.05 level.

Normal probability plots of the samples show that the normality assumption is justified. Figures 5.8 and 5.9 contain the plots for $\psi=0.1$ and $\psi=10$, respectively. The points seem to lie on a straight line. Plots for other

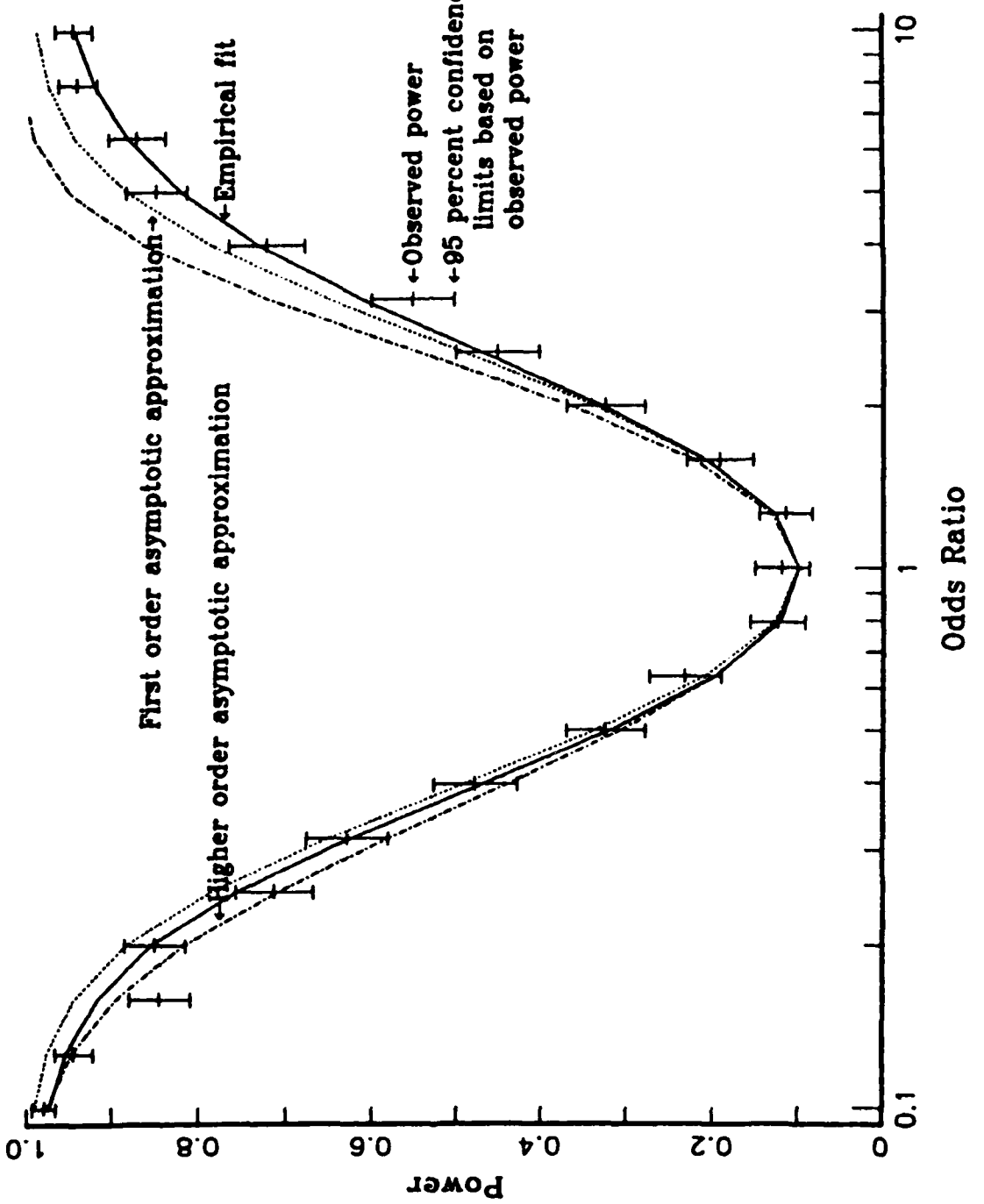

Power as a Function of Odds Ratio
First order asymptotic approximation
Higher order asymptotic approximation
Empirical fit
Observed power
95 percent confidence limits based on observed power
Power
Odds Ratio
0
0.2
0.4
0.6
0.8
1.0
0.1
1
10

Figure 5.7.

Figure 5.8. Normal Probability Plot for Score Statistics

Generated with $\psi = 0.10$

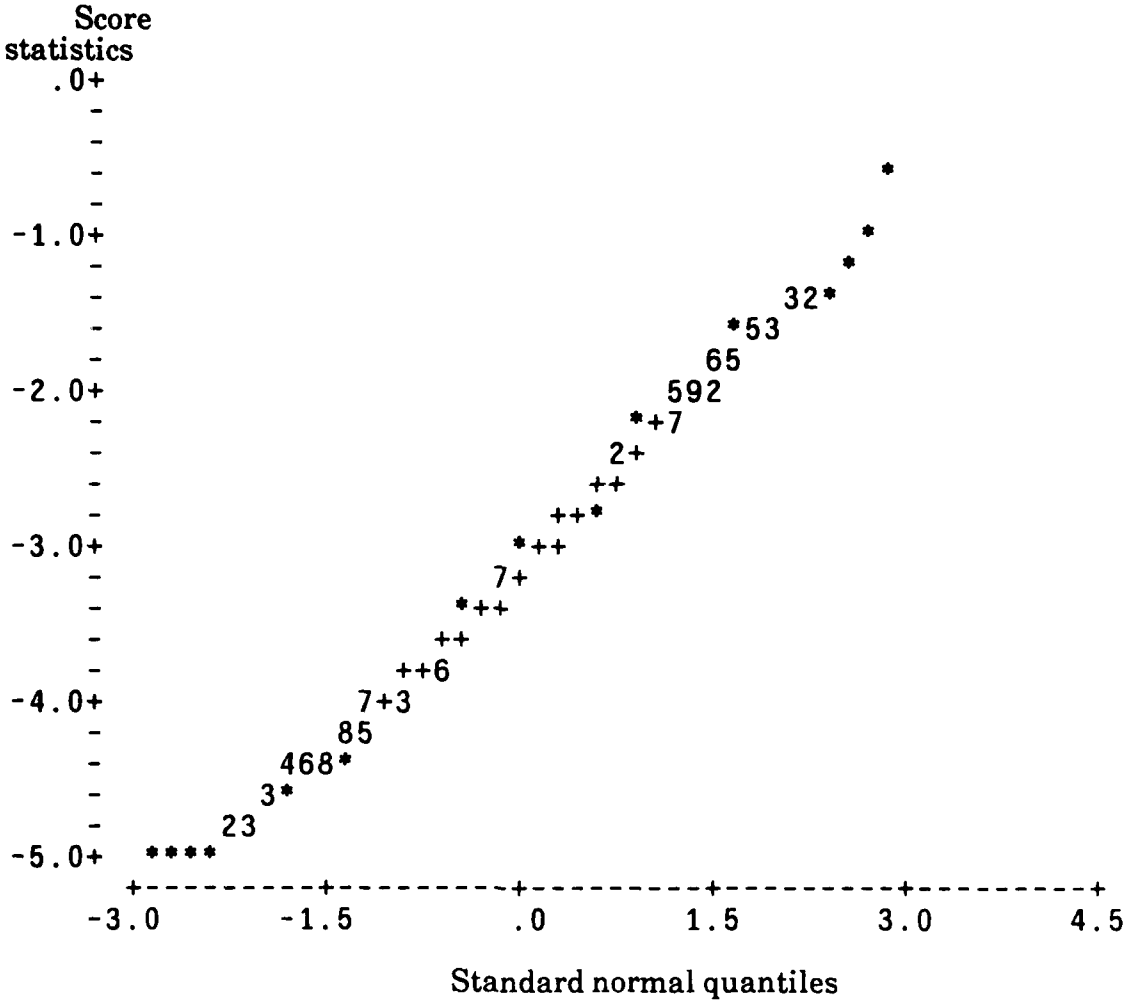
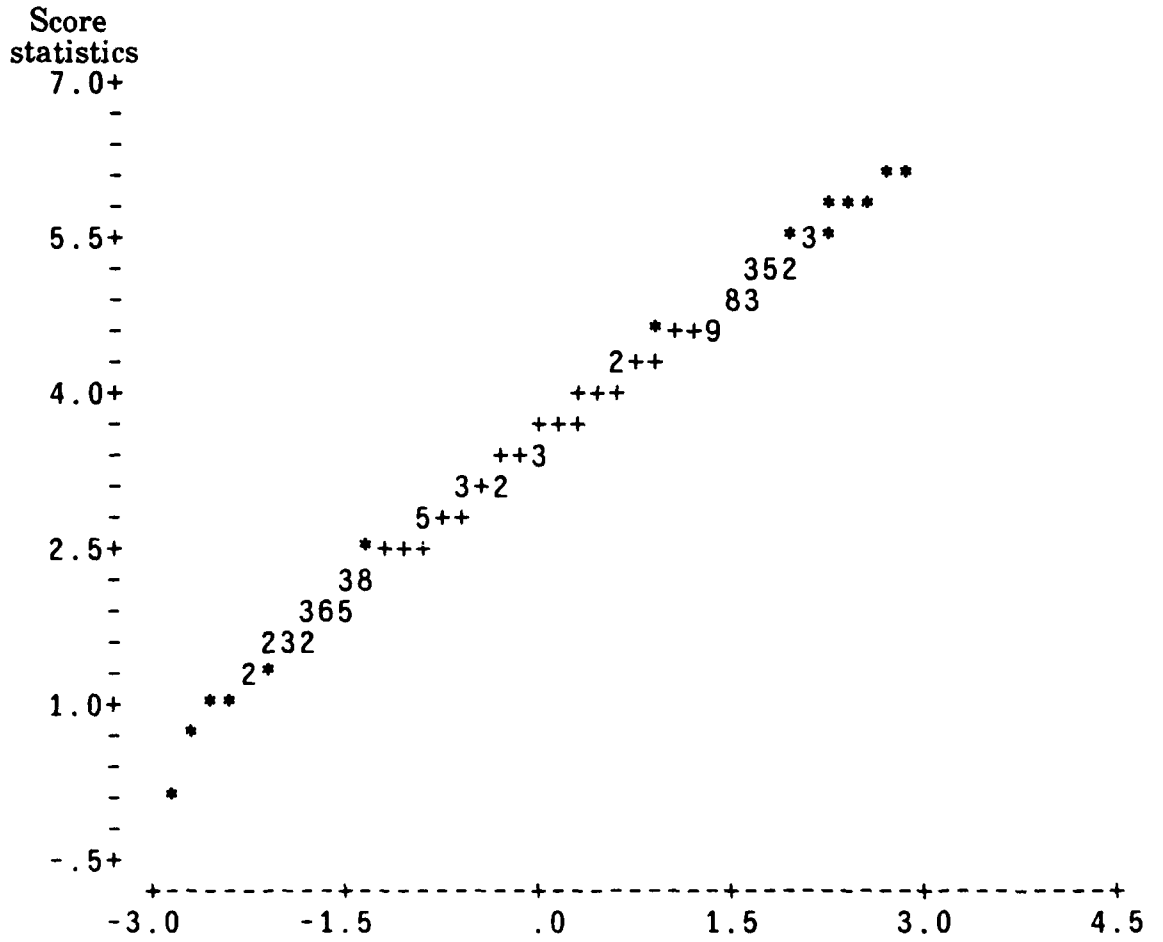


Figure 5.9. Normal Probability Plot for Score Statistics

Generated with $\psi = 10$



values of the odds ratio are similar.

To see if the empirical fit holds more generally, I performed a second simulation. I generated $\{X_t\}$ via the relation $X_t = \rho X_{t-1} + u_t$, where $\{u_t\}$ is a sequence of independent standard normal random variables, $\rho = 0.5$, and $X_0 = 0$. The sample size, intercept, and slope were 150, -1, and 0.5, respectively.

The results are summarized in Figures 5.10-5.12. In Figure 5.10 the dotted line is the curve

$$\mu = 5.39^{1/2}(\log \varphi) - 0.065\ (5.39/3.10)^{1/2}(\log \varphi)^3,$$

where 5.39 is the information number in the second simulation and 3.10 is the information number in the original simulation. The asymptotic line gives an excellent fit to the observed means when $\varphi \geq 0.5$, unlike the previous case. For $\varphi < 0.5$ the behavior is qualitatively similar to that found in the first simulation, but the empirical fit is not as good.

In Figure 5.11 the results are similar to those found earlier, but there seems to be a steeper slope for $\varphi < 1$ and there is some indication that when $\varphi > 1$ the standard deviation exceeds 1.

Figure 5.12 contains a plot of the observed power, first order asymptotic power, and the power obtained by the empirical fit. The fitted power may be marginally better than the asymptotic power for $\varphi < 0.6$, but the asymptotic power seems better elsewhere.

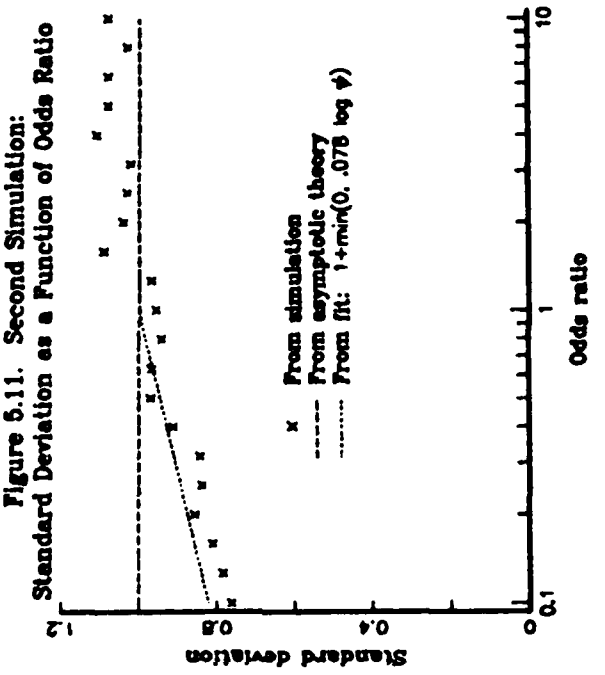


Figure 5.11. Second Simulation:
Standard Deviation as a Function of Odds Ratio

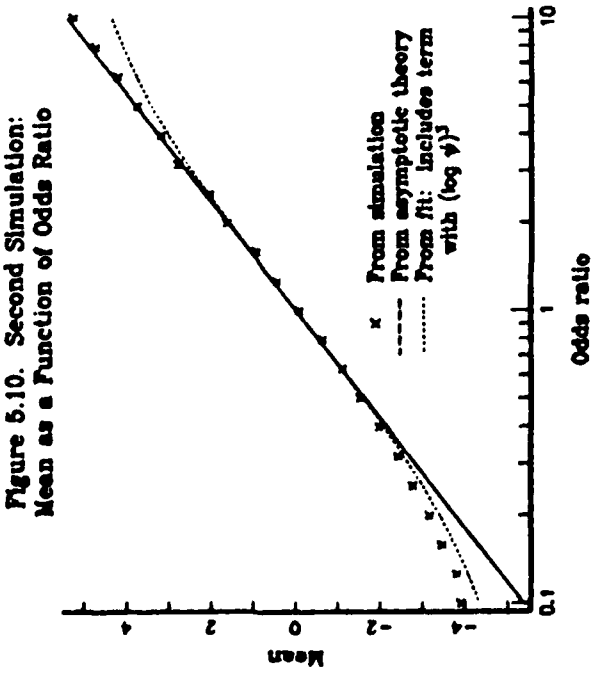


Figure 5.10. Second Simulation:
Mean as a Function of Odds Ratio

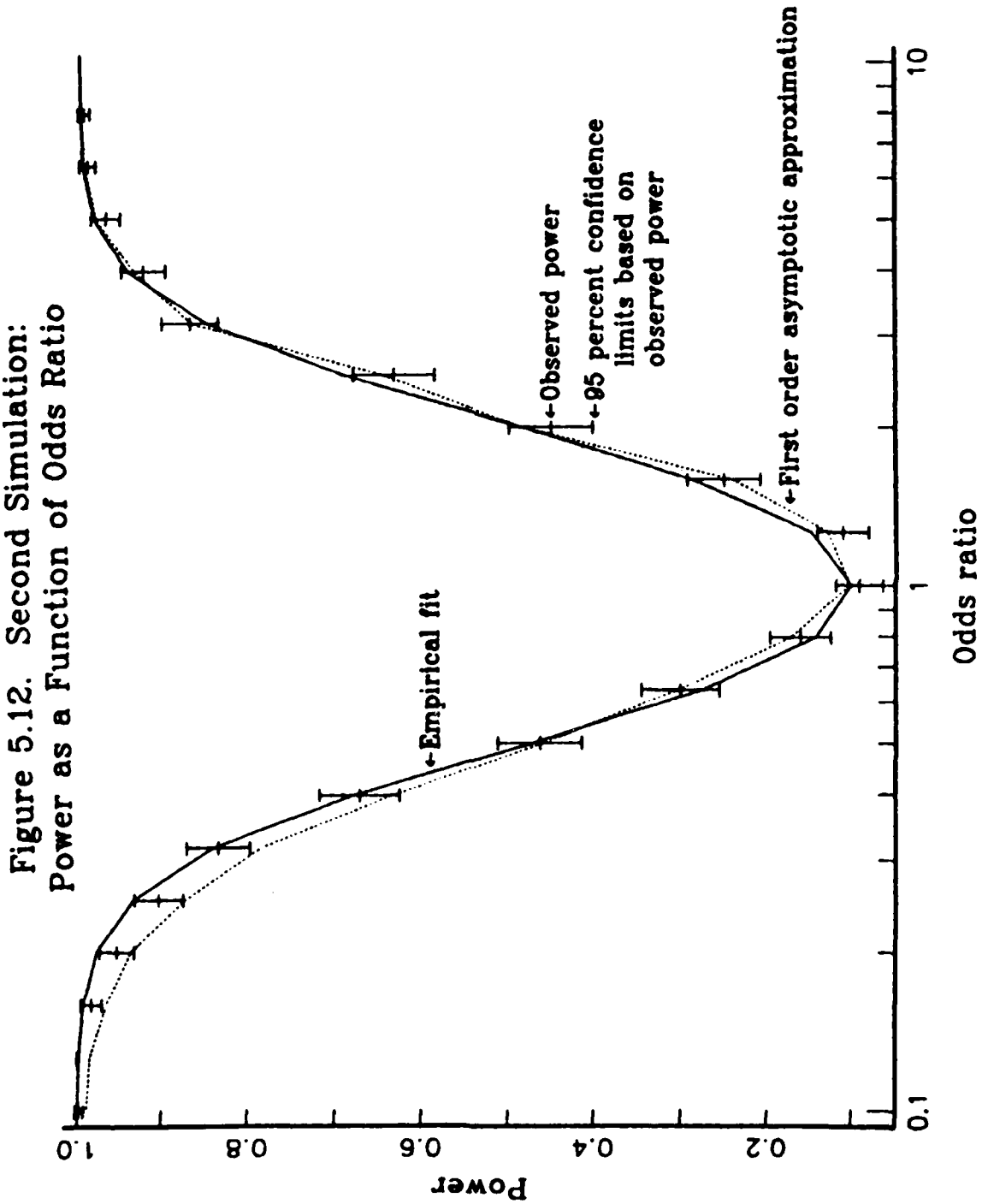


Figure 5.12. Second Simulation: Power as a Function of Odds Ratio

In summary, the simple first order approximation to the distribution of the score statistic gives a power function that is not significantly different from the power observed in a random sample when the odds ratio differs from one by a factor of no more than two or three. For more extreme odds ratios the approximation overestimates the power, though it might be considered adequate as a qualitative description of the power. A higher order approximation does not significantly improve the agreement between the empirical and theoretical power curves.

A look at the sample of score statistics, though, shows that their distribution is normal but with parameters that vary systematically from those predicted in the asymptotic approximation. Fitting a smooth curve to these parameters allows calculation of a power function more in agreement with the observed power. This empirical fit can be obtained by simulation for any given set of X's, but a fit obtained for one particular set of X's does not appear to be valid for other X's.

Chapter 6

Missing Data

In this chapter I will examine the effect of missing Y values on the score statistic and to a more limited extent on the estimation process. I will assume that the corresponding X values are not missing, and I will comment on this assumption where appropriate.

6.1 Effect on the score test for independence

In this chapter I will examine the effect of missing Y values on the score statistic for testing independence (ψ=1) and on the estimation process. Recall that with no missing data the log likelihood function and its derivatives can be written

$$L = \log P[Y_1=y_1] + \sum Y_t \log \pi + (1-Y_t) \log (1-\pi),$$

$$\frac{\partial L}{\partial \psi} = \sum \frac{Y-\pi}{\pi(1-\pi)} \frac{\partial \pi}{\partial \psi},$$

$$\frac{\partial L}{\partial \beta} = \sum \frac{Y-\pi}{\pi(1-\pi)} \frac{\partial \pi}{\partial \beta},$$

where $\pi = P[Y_t=1|Y_{t-1}]$. This leads to a score function for ψ that contains products of consecutive Y values.

Suppose Y_{t-1} is not observed but X_{t-1}, Y_{t-2}, and X_{t-2} are observed. If I define $\pi_{t,m} = P[Y_t=1|Y_{t-m}]$ and $\pi_t(z) = P[Y_t=1|Y_{t-1}=z)$, then if $Y_t=1$ the contribution to the log likelihood from Y_t is

$$\begin{aligned}\log \pi_{t,2} &= \log(P[Y_t=1|Y_{t-1}=1]\ P[Y_{t-1}=1|Y_{t-2}] \\ &\qquad + P[Y_t=1|Y_{t-1}=0]\ P[Y_{t-1}=0|Y_{t-2}]) \\ &= \log(\pi_t(1)\ \pi_{t-1}(Y_{t-2}) + \pi_t(0)\ (1-\pi_{t-1}(Y_{t-2}))),\end{aligned}$$

and if $Y_t=0$ the contribution is $\log(1-\pi_{t,2})$. (Note that these are random variables; they depend on Y_{t-2}.)

Using this expression for the likelihood, the score test can still be performed and parameter estimates can still be found by maximum likelihood when a single Y value is missing. If two or more consecutive Y values are missing, $\pi_{t,m}$ can be written similarly by summing over all possible values of Y_{t-1} through Y_{t-m+1}. This cannot be done, however, if the corresponding X's are missing, because the likelihood is a function of all these X's.

Suppose first that only Y_{t-1} is missing. The contribution of the t<u>th</u> term of the likelihood to the score function for φ can be written

$$Y_t \frac{1}{\pi_{t,2}} \frac{\partial \pi_{t,2}}{\partial \varphi} - (1-Y_t) \frac{1}{1-\pi_{t,2}} \frac{\partial \pi_{t,2}}{\partial \varphi},$$

with

$$\frac{\partial \pi_{t,2}}{\partial \varphi} = \frac{\partial \pi_t(1)}{\partial \varphi}\ \pi_{t-1}(Y_{t-2}) + \pi_t(1)\ \frac{\partial \pi_{t-1}(Y_{t-2})}{\partial \varphi}$$

$$+ \frac{\partial \pi_t(0)}{\partial \varphi}\ (1-\pi_{t-1}(Y_{t-2})) - \pi_t(0)\ \frac{\partial \pi_{t-1}(Y_{t-2})}{\partial \varphi}\ .$$

Evaluated at $\varphi=1$ this simplifies to

$$\frac{\partial \pi_{t,2}}{\partial \varphi} = [(1-p_{t-1})p_t(1-p_t)]\ p_{t-1} + p_t\ [(Y_{t-2}-p_{t-2})p_{t-1}(1-p_{t-1})]$$

$$+ [-p_{t-1}p_t(1-p_t)]\ (1-p_{t-1}) - p_t\ [(Y_{t-2}-p_{t-2})p_{t-1}(1-p_{t-1})]$$

$$= 0.$$

Therefore there is no contribution to the score statitistic due to the dependence between Y_t and Y_{t-2} if Y_{t-1} is missing.

This is also true if more than one observation is missing, as can be seen by writing

$$\pi_{t,k} = \pi_{t,2}(1)\ \pi_{t-2,k-2}(Y_{t-k}) + \pi_{t,2}(0)\ (1-\pi_{t-2,k-2}(Y_{t-k}))\ ,$$

so that

$$\frac{\partial \pi_{t,k}}{\partial \varphi} = \frac{\partial \pi_{t,2}(1)}{\partial \varphi}\ \pi_{t-2,k-2}(Y_{t-k}) + \pi_{t,2}(1)\ \frac{\partial \pi_{t-2,k-2}(Y_{t-k})}{\partial \varphi}$$

$$+ \frac{\partial \pi_{t,2}(0)}{\partial \varphi}\ (1-\pi_{t-2,k-2}(Y_{t-k})) - \pi_{t,2}(0)\ \frac{\partial \pi_{t-2,k-2}(Y_{t-k})}{\partial \varphi}\ .$$

The first factors in the first and third terms are zero (from above), while the second and fourth terms cancel, since $\pi_{t,2}(1) = \pi_{t,2}(0)$ under the null hypothesis.

In summary, if Y_{t-1} is missing there is no contribution to the score statistic from the dependence between Y_{t-2} and Y_t. The only positive terms in the score statistic are those that involve consecutive observations. The reduced dependence between Y_{t-k} and Y_t cannot be measured by this statistic for any $k>1$.

If, for example, every second observation in a sample were missing, it might be possible to reparametrize and use $\theta = f(\varphi)$ as the measure of dependence, where f is such that $f'(1)=0$. The score function would be $(\partial L/\partial\varphi)/f'(\varphi)$, so it might approach a finite non-zero value as φ approaches 1 for a suitably chosen f. However in such a case it may be simpler to assume a model in which the odds ratio between Y_t and Y_{t-2} is constant for all t, and to treat the problem as if no data were missing. This is a different model, but it might be nearly the same if the marginal probabilities are nearly constant.

In the more likely case of some consecutive observations together with some separated by gaps, however, the contributions from consecutive observations are infinitely larger than the contributions from observations separated by gaps. Therefore no such reparametrization is possible in this case.

There is no standard procedure for modifying the Durbin-Watson statistic for missing values when testing for serial correlation in a least squares regression. Three possible modifications are given by Savin and White

(1978) and by Honohan and McCarthy (1982). Two of these are similar to the score statistic presented here; they omit the terms in the statistic that involve missing values. The other simply removes any missing values and treats any surrounding observations as if they were taken at consecutive times.

6.2 Effect on estimation of the odds ratio

It is still possible to write down the likelihood function in the presence of missing data, so the maximum likelihood estimates can be found. Because there is no simple general expression for the derivatives of the log likelihood, however, it would probably be easier to use a derivative-free maximization procedure in this case. Using Newton's method would require calculating and programming first and second derivatives for every "gap length" observed in the sample.

It seems that some information would be lost if n observations span m>n time periods, in comparison with the information in n consecutive observations. I will show below that this is usually, but not always, the case.

Suppose the coefficients and therefore the marginal probabilities $\{p_t\}$ are known, and I want to estimate the odds ratio ψ. Then the contribution to the Fisher information for estimating ψ can be calculated for an observation both when the previous value is observed and when it is missing. The ratio of the two information numbers gives the asymptotic relative efficiency of the two estimators and provides a measure of the information loss caused by the intervening missing observation. These information

numbers are calculated below.

Table 6.1 contains this ratio under the following conditions. In the case of consective observations I assume they each have marginal probability 0.5 of "success." In the second case I assume the same two random variables are separated by a single missing value whose marginal probability of success takes values between 0.5 and 0.95 in increments of 0.05. In both cases I use odds ratios increasing from 1/32 to 32 in multiples of 2.

The surprising feature in this table is the appearance of ratios larger than 1 for p_{t-1} near 0.5 and for extreme values of the odds ratio. This indicates that for those parameter values, if only two observations can be taken it is advantageous not to take them consecutively, but to allow an intervening value to pass unobserved. When ψ is very large two consecutive observations take the same value with a very high probability, so little information is gained about the value of ψ. Skipping an observation reduces the dependence and provides more information.

This effect is more pronounced in Table 6.2, where the marginal probabilities of the observed values are 0.1. Here the ratio exceeds 1 for certain values of the other marginal probabilities when the odds ratio is as high as 0.25. For these values of the odds ratio, however, the dependence is not simply reduced by inserting a missing observation, its direction is also changed.

Suppose a statistician is able to take a fixed number of observations of a

Table 6.1. Asymptotic Relative Efficiency

when $p_t = 0.5$

Asymptotic relative efficiency of the maximum likelihood estimator of the odds ratio when a single missing value intervenes between two observations, as compared with two consecutive observations. P is the marginal probability that the missing variable is 1. The observed variables all have marginal probability 0.5 of success.

	Odds Ratio					
P	1	2	4	8	16	32
.50	.000	.114	.400	.743	1.059	1.314
.55	.000	.112	.388	.708	.974	1.134
.60	.000	.104	.353	.612	.767	.761
.65	.000	.093	.301	.482	.529	.434
.70	.000	.078	.239	.346	.328	.225
.75	.000	.061	.174	.225	.185	.109
.80	.000	.043	.114	.131	.094	.049
.85	.000	.026	.065	.066	.042	.020
.90	.000	.013	.028	.026	.015	.007
.95	.000	.003	.007	.006	.003	.001

P	1	1/2	1/4	1/8	1/16	1/32
.50	.000	.114	.400	.743	1.059	1.314
.55	.000	.112	.388	.708	.974	1.134
.60	.000	.104	.353	.612	.767	.761
.65	.000	.093	.301	.482	.529	.434
.70	.000	.078	.239	.346	.328	.225
.75	.000	.061	.174	.225	.185	.109
.80	.000	.043	.114	.131	.094	.049
.85	.000	.026	.065	.066	.042	.020
.90	.000	.013	.028	.026	.015	.007
.95	.000	.003	.007	.006	.003	.001

Table 6.2. Asymptotic Relative Efficiency

when $p_t = 0.1$

Asymptotic relative efficiency of the maximum likelihood estimator of the odds ratio when a single missing value intervenes between two observations, as compared with two consecutive observations. P is the marginal probability that the missing variable is 1. The observed variables all have marginal probability 0.1 of success.

	Odds Ratio					
P	1	2	4	8	16	32
.50	.000	.061	.085	.054	.024	.009
.55	.000	.051	.062	.036	.015	.005
.60	.000	.041	.044	.023	.009	.003
.65	.000	.031	.030	.015	.005	.002
.70	.000	.023	.020	.009	.003	.001
.75	.000	.016	.013	.005	.002	.001
.80	.000	.010	.007	.003	.001	.000
.85	.000	.006	.004	.001	.000	.000
.90	.000	.002	.001	.001	.000	.000
.95	.000	.001	.000	.000	.000	.000

P	1	1/2	1/4	1/8	1/16	1/32
.50	.000	.146	.464	.613	.523	.350
.55	.000	.169	.618	.915	.851	.601
.60	.000	.187	.800	1.351	1.394	1.057
.65	.000	.200	1.005	1.971	2.309	1.925
.70	.000	.203	1.218	2.822	3.864	3.664
.75	.000	.194	1.404	3.919	6.482	7.336
.80	.000	.179	1.505	5.140	10.595	15.241
.85	.000	.131	1.434	6.040	15.678	30.099
.90	.000	.079	1.095	5.668	17.669	41.728
.95	.000	.027	.480	3.033	10.030	21.221

binary first-order Markov process with known marginal probabilities, and his object is to estimate the odds ratio of the matrix of transition probabilities. The calculations used in creating these tables could be used to determine an optimal sampling scheme. For example, if the marginal probabilities were all 0.5, the first row in Table 6.1 indicates that more information about the odds ratio is obtained by observing the process at times 1 and 3 than at times 1 and 2 if $\varphi \geq 16$.

These tables show that for certain sequences of marginal probabilities, if the odds ratio is known a priori to lie in a certain region it may be profitable to observe the process intermittently rather than continuously. In such cases calculation of the above ratio could allow an optimal sampling scheme to minimize the asymptotic variance of the estimate of the odds ratio. However for most values of the odds ratio and marginal probabilities it is better to take consecutive observations. In these tables and in all others I examined, the ratio is always less than 0.21 for odds ratios within a factor of 2 of 1. It is unlikely that in any realistic application the evidence presented here warrants letting values pass unobserved.

This phenomenon was observed previously in the case of stationary first order autoregressive processes with known mean 0 by Dunsmuir (1981). (His univariate continuous model is analagous to the binary process with known constant marginal probabilities.) His results can be used to show that for values of the autoregressive parameter larger than $3^{-1/2}$, it is more advantageous to skip a time point between observations than to take two

consecutive observations of the process.

Dunsmuir goes further and derives formulas that can be used to compute the asymptotic relative efficiencies whenever the frequencies of gap lengths are given. He applies these results to two types of sampling schemes: Bernoulli sampling, where the series is observed or not at each time point according to a sequence of independent Bernoulli random variables, and regular A-B sampling, where the series is observed according to a repeating pattern of A observations and B missing values.

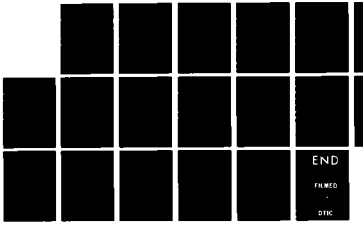
In the binary case the asymptotic relative efficiency could be calculated for any particular pattern of regular A-B sampling. Bernoulli sampling, on the other hand, involves random unbounded stretches of missing observations, and since no general expression for the information as a function of gap length seems to be feasible, the asymptotic relative efficiency cannot be calculated in closed form.

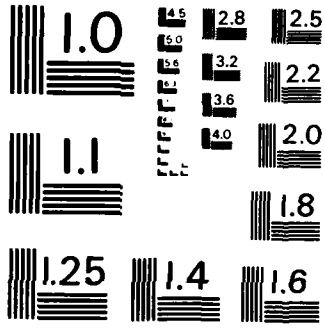
In the remainder of this section I will calculate the Fisher information only for the two cases used in computing the ratios in Table 1: consecutive observations and observations separated by a single missing value. The Fisher information for consecutive observations is given in Chapter 4, but I repeat it here for convenience.

For an observation Y_2 preceded by another observation Y_1, the contributions to the log likelihood and its derivatives are

AD-A161 274 A MODEL FOR SERIAL DEPENDENCE IN LOGISTIC REGRESSION (U) MASSACHUSETTS INST OF TECH CAMBRIDGE STATISTICS CENTER T P LANE SEP 85 TR-38-ONR N00014-75-C-0555 2/2

UNCLASSIFIED F/G 12/1 NL





$$L = Y_1Y_2 \log a_2 + (1-Y_1)Y_2 \log (p_2-a_2) + Y_1(1-Y_2) \log (p_1-a_2)$$
$$+ (1-Y_1)(1-Y_2) \log (1+a_2-p_1-p_2) ,$$

$$\frac{\partial L}{\partial \varphi} = \frac{\partial a_2}{\partial \varphi} \left[\frac{Y_1Y_2}{a_2} - \frac{(1-Y_1)Y_2}{p_2-a_2} - \frac{Y_1(1-Y_2)}{p_1-a_2} + \frac{(1-Y_1)(1-Y_2)}{1+a_2-p_1-p_2} \right] ,$$

$$\frac{\partial^2 L}{\partial \varphi^2} = \frac{\partial^2 a_2}{\partial \varphi^2} \left[\frac{Y_1Y_2}{a_2} - \frac{(1-Y_1)Y_2}{p_2-a_2} - \frac{Y_1(1-Y_2)}{p_1-a_2} + \frac{(1-Y_1)(1-Y_2)}{1+a_2-p_1-p_2} \right]$$
$$- \left[\frac{\partial a_2}{\partial \varphi} \right]^2 \left[\frac{Y_1Y_2}{a_2^2} + \frac{(1-Y_1)Y_2}{(p_2-a_2)^2} + \frac{Y_1(1-Y_2)}{(p_1-a_2)^2} + \frac{(1-Y_1)(1-Y_2)}{(1+a_2-p_1-p_2)^2} \right] .$$

The expectation of the first term in the second derivative is 0, so the Fisher information is

$$- E \left[\frac{\partial^2 L}{\partial \varphi^2} \right] = \left[\frac{\partial a_2}{\partial \varphi} \right]^2 \left[\frac{1}{a_2} + \frac{1}{p_2-a_2} + \frac{1}{p_1-a_2} + \frac{1}{1+a_2-p_1-p_2} \right]$$

$$= \left[\frac{(p_1-a_2)(p_2-a_2)}{1+(\varphi-1)(p_1+p_2-2a_2)} \right]^2 \left[\frac{1}{a_2} + \frac{1}{p_2-a_2} + \frac{1}{p_1-a_2} + \frac{1}{1+a_2-p_1-p_2} \right].$$

This is the information obtained from an observation when the preceding observation is not missing.

Now suppose a single unobserved Y_2 intervenes between two observed values Y_1 and Y_3. For convenience define the following four quantities and their derivatives:

$$D_1 = (1-p_2)a_2a_3 + p_2(p_1-a_2)(p_3-a_3)$$

$$D_2 = (1-p_2)(p_2-a_2)a_3 + p_2(1+a_2-p_1-p_2)(p_3-a_3)$$

$$D_3 = (1-p_2)a_2(p_2-a_3) + p_2(p_1-a_2)(1+a_3-p_2-p_3)$$

$$D_4 = (1-p_2)(p_2-a_2)(p_2-a_3) + p_2(1+a_2-p_1-p_2)(1+a_3-p_2-p_3)$$

$$\frac{\partial D_1}{\partial \varphi} = (1-p_2)\left[\frac{\partial a_2}{\partial \varphi} a_3 + a_2 \frac{\partial a_3}{\partial \varphi}\right]$$

$$+ p_2\left[-\frac{\partial a_2}{\partial \varphi}(p_3-a_3) - (p_1-a_2)\frac{\partial a_3}{\partial \varphi}\right]$$

$$\frac{\partial D_2}{\partial \varphi} = (1-p_2)\left[-\frac{\partial a_2}{\partial \varphi} a_3 + (p_2-a_2)\frac{\partial a_3}{\partial \varphi}\right]$$

$$+ p_2\left[\frac{\partial a_2}{\partial \varphi}(p_3-a_3) - (1+a_2-p_1-p_2)\frac{\partial a_3}{\partial \varphi}\right]$$

$$\frac{\partial D_3}{\partial \varphi} = (1-p_2)\left[\frac{\partial a_2}{\partial \varphi}(p_2-a_3) - a_2 \frac{\partial a_3}{\partial \varphi}\right]$$

$$+ p_2\left[-\frac{\partial a_2}{\partial \varphi}(1+a_3-p_2-p_3) + (p_1-a_2)\frac{\partial a_3}{\partial \varphi}\right]$$

$$\frac{\partial D_4}{\partial \varphi} = (1-p_2)\left[-\frac{\partial a_2}{\partial \varphi}(p_2-a_3) - (p_2-a_2)\frac{\partial a_3}{\partial \varphi}\right]$$

$$+ p_2\left[\frac{\partial a_2}{\partial \varphi}(1+a_3-p_2-p_3) + (1+a_2-p_1-p_2)\frac{\partial a_3}{\partial \varphi}\right].$$

Simplifying these derivatives leads to

$$\frac{\partial D_1}{\partial \varphi} = -\frac{\partial D_2}{\partial \varphi} = -\frac{\partial D_3}{\partial \varphi} = \frac{\partial D_4}{\partial \varphi} = (\alpha_3 - p_2 p_3)\frac{\partial \alpha_2}{\partial \varphi} + (\alpha_2 - p_1 p_2)\frac{\partial \alpha_3}{\partial \varphi} .$$

Then letting $A = p_2(1-p_2)$, the contribution to the likelihood from the Y_3 term can be written

$$\log L = Y_1 Y_3 \log [\, D_1/A \,] + (1-Y_1)Y_3 \log [\, D_2/A \,]$$

$$+ Y_1(1-Y_3) \log [\, D_3/A \,] + (1-Y_1)(1-Y_3) \log [\, D_4/A \,] .$$

The score function is therefore

$$\frac{\partial \log L}{\partial \varphi} = \frac{Y_1 Y_3}{D_1}\frac{\partial D_1}{\partial \varphi} + \frac{(1-Y_1)Y_3}{D_2}\frac{\partial D_2}{\partial \varphi} + \frac{Y_1(1-Y_3)}{D_3}\frac{\partial D_3}{\partial \varphi}$$

$$+ \frac{(1-Y_1)(1-Y_3)}{D_4}\frac{\partial D_4}{\partial \varphi}$$

$$= \left[\frac{\partial \alpha_2}{\partial \varphi}(\alpha_3 - p_2 p_3) + \frac{\partial \alpha_3}{\partial \varphi}(\alpha_2 - p_1 p_2) \right]$$

$$\times \left[\frac{Y_1 Y_3}{D_1} - \frac{(1-Y_1)Y_3}{D_2} - \frac{Y_1(1-Y_3)}{D_3} + \frac{(1-Y_1)(1-Y_3)}{D_4} \right] .$$

The expectation of the second factor in square brackets is zero, so the Fisher information does not contain the derivative of the first term. The expectation of the second derivative is

$$E\left[\frac{\partial^2 \log L}{\partial \varphi^2}\right] = \left[\frac{\partial \alpha_2}{\partial \varphi}(\alpha_3 - p_2 p_3) + \frac{\partial \alpha_3}{\partial \varphi}(\alpha_2 - p_1 p_2)\right]$$

$$\times E\left[-\frac{Y_1 Y_3}{D_1^2}\frac{\partial D_1}{\partial \varphi} + \frac{(1-Y_1)Y_3}{D_2^2}\frac{\partial D_2}{\partial \varphi} + \frac{Y_1(1-Y_3)}{D_3^2}\frac{\partial D_3}{\partial \varphi}\right.$$

$$\left. - \frac{(1-Y_1)(1-Y_3)}{D_4^2}\frac{\partial D_4}{\partial \varphi}\right]$$

$$= \left[\frac{\partial \alpha_2}{\partial \varphi}(\alpha_3 - p_2 p_3) + \frac{\partial \alpha_3}{\partial \varphi}(\alpha_2 - p_1 p_2)\right] \frac{1}{p_2(1-p_2)}$$

$$\times \left[-\frac{1}{D_1}\frac{\partial D_1}{\partial \varphi} + \frac{1}{D_2}\frac{\partial D_2}{\partial \varphi} + \frac{1}{D_3}\frac{\partial D_3}{\partial \varphi} - \frac{1}{D_4}\frac{\partial D_4}{\partial \varphi}\right]$$

$$= \left[\frac{\partial \alpha_2}{\partial \varphi}(\alpha_3 - p_2 p_3) + \frac{\partial \alpha_3}{\partial \varphi}(\alpha_2 - p_1 p_2)\right]^2 \frac{-1}{p_2(1-p_2)}$$

$$\times \left[\frac{1}{D_1} + \frac{1}{D_2} + \frac{1}{D_3} + \frac{1}{D_4}\right].$$

The negative of the latter quantity is the expression for the Fisher information that was used in computing Tables 6.1 and 6.2.

Chapter 7

Graphics

Autocorrelation in least squares regression is often easily detected in a plot. If the coefficients are estimated by least squares and the estimates are used to compute residuals $\{r_t\}$, then a plot of r_t as a function of t will generally show any serial dependence

Similar plots are not as useful in the serial dependence model. One reason for this is that the residuals in some sense cannot be separated from the fitted values as they can in least squares.

The two models differ in the constraints placed on the residuals. In least squares, adding any fitted value $\hat{y}_t$ to any residual r_t produces an acceptable observation $y_t = \hat{y}_t + r_t$. In binary regression for each fitted probability $\hat{p}_t$ there are only two possible residuals, $1-\hat{p}_t$ and $-\hat{p}_t$.

Another difference is in the effect on joint probabilities of the marginal probabilities. In least squares, if the true error process is $\{\varepsilon_t\}$, the probability that ε_t and ε_{t-1} have the same sign is a function only of the autocorrelation. In binary regression the probability that y_t-p_t and $y_{t-1}-p_{t-1}$ have the same sign depends not only on the odds ratio but also on p_t and p_{t-1}. If both marginal probabilities are close to one, then for some ρ values smaller than 1 the probability that the two errors have the

same sign is larger than 0.5.

Figure 7.1 contains three plots of residuals from an ordinary logistic regression as a function of time. In each case there are 100 observations with marginal probabilities satisfying $\log (p_t/(1-p_t)) = 1+x_t$, where $\{x_t\}$ are independent standard normal random variables. I chose these plots subjectively as typical results for sequences generated using $\psi = 4$, 1, and 0.25. The values of the score statistics for testing independence are also given on the plot.

The residuals used in the plot are standardized as follows:

$$r_t = n^{1/4} \frac{y_t - \hat{p}_t}{[\sum \hat{p}_t(1-\hat{p}_t)\hat{p}_{t-1}(1-\hat{p}_{t-1})]^{1/4}}$$

I chose this scale because $n^{-1/2} r_t r_{t-1}$ is a component of the score statistic. This will be more important in Figure 7.2. Here it does not change the visual impression of the plots, only the scale markings on the vertical axis.

At first glance the appearance of each of these plots resembles that of the least squares residuals from a process with negative autocorrelation, since the lines are jagged and they cross the axis frequently. However this is in large part due to the discreteness of the residuals rather than their serial dependence.

If the three plots are compared, some features become apparent. Wide hills and valleys are more common for large values of the odds ratio than

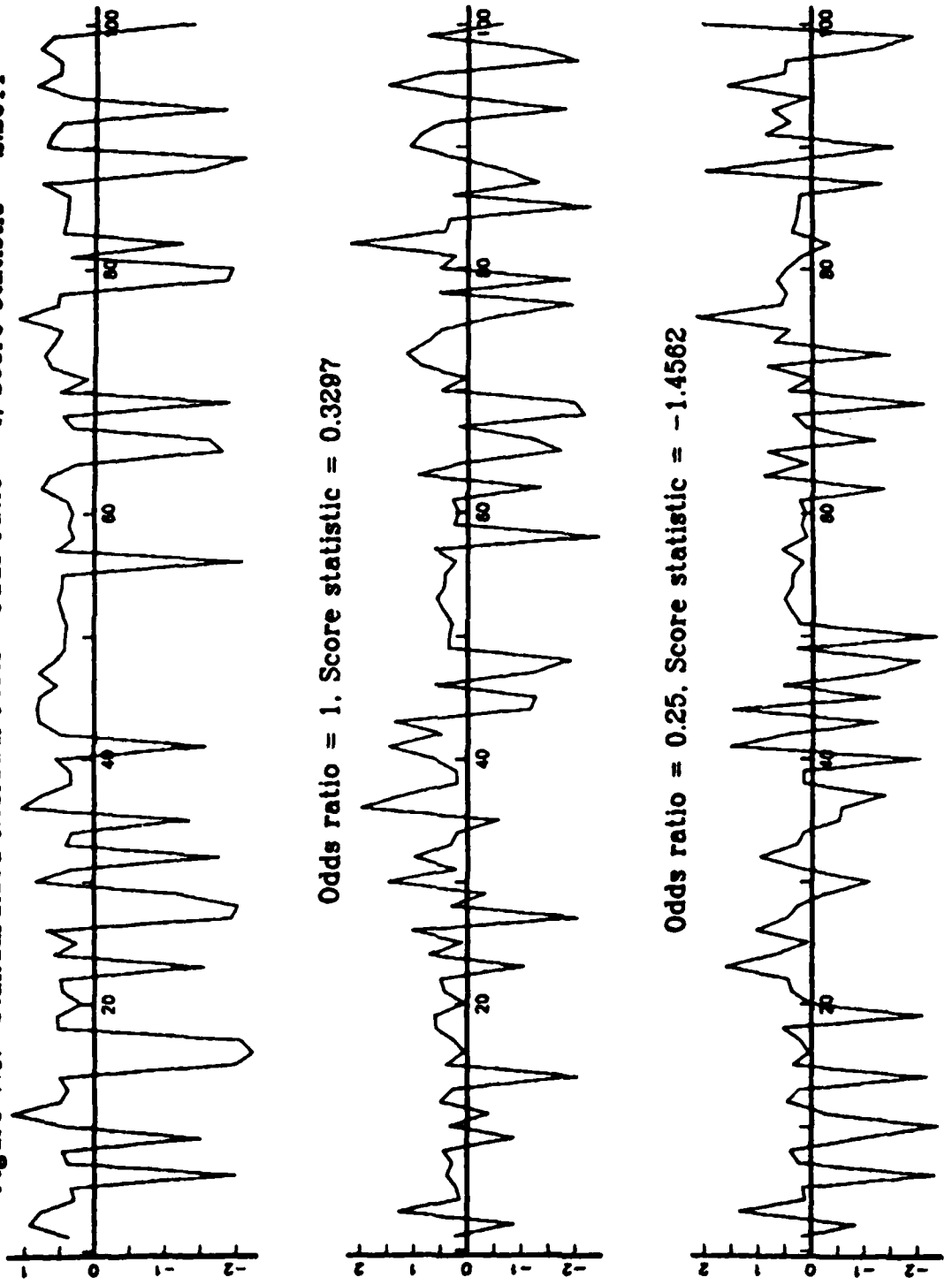


Figure 7.1. Standardized Residual Plots.

for small values. This is especially true for deep valleys; consecutive large negative residuals occasionally appear when φ=4 (such as at t=16 and 28) but rarely when φ=0.25. Also more numerous peaks and valleys for low values of φ give the impression that the plots are more "stretched" horizontally moving from the bottom plot to the top.

Unfortunately none of these visual impressions is as striking as the values of the score statistics associated with each plot.

Because the score statistic is so useful, it can be presented visually by noting

$$W^{1/2} = \sum_{t=2}^{n} \frac{y_t - \hat{p}_t}{[\sum \hat{p}_t(1-\hat{p}_t)\hat{p}_{t-1}(1-\hat{p}_{t-1})]^{1/4}} \frac{y_{t-1} - \hat{p}_{t-1}}{[\sum \hat{p}_t(1-\hat{p}_t)\hat{p}_{t-1}(1-\hat{p}_{t-1})]^{1/4}}$$

$$= n^{-1/2} \sum r_t r_{t-1} ,$$

so a plot of r_t against r_{t-1} may be useful. Figure 7.2 contains these plots for the same residual vectors used in Figure 7.1.

There are several features in these plots that are worth exploring. First, the plot for φ=4 appears to be divided into four clusters. This phenomenon occasionally appears in this type of plot for any value of φ; it is caused more by the marginal probabilities than by the value of φ. If there are few small marginal probabilities, then there will be few small negative residuals. Therefore the scatter plot will be sparse just below the horizontal axis and just to the left of the vertical axis.

A second feature is the presence of many points near the origin in the

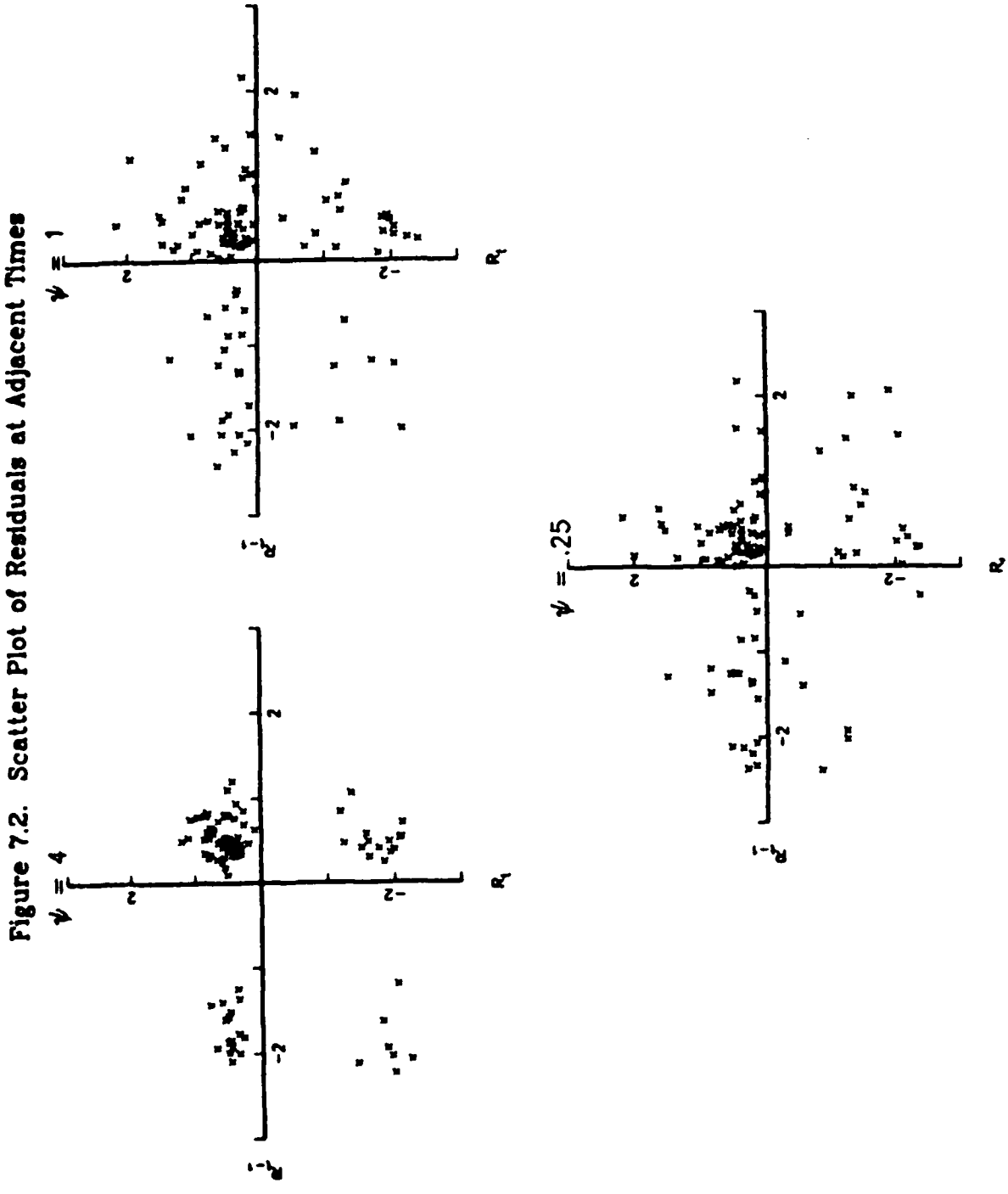


Figure 7.2. Scatter Plot of Residuals at Adjacent Times

first quadrant. This again depends on the marginal probabilities; if there are several consecutive marginal probabilities near 1, there will be several small positive residuals. If the odds ratio is large, though, this tendency will be more pronounced, as in Figure 7.2.

A third feature, perhaps not as obvious as the other two, is that there are more extreme points in quadrants I and III if $\psi>1$ and more in quadrants II and IV if $\psi<1$. (Here I define "extreme" points as those with both coordinates far from 0.) It is this feature that seems to be a good indicator of serial dependence.

Extreme points indicate that two consecutive observations took values for which the marginal probabilities were relatively low. The odds ratio, however, is a measure of how the joint probability differs from the product of the joint probabilities, so the frequency of these events gives some information about the odds ratio. (In Figure 7.2 there are no extreme points in quadrant I when $\psi=4$, but this is not typical of such plots.)

The score statistic is a multiple of a sum of the products of consecutive standardized residuals, so the contribution of a point depends on the product of its coordinates. Therefore in interpreting these plots it is helpful to consider each point in relation to the hyperbolas defined by constant values of $r_t r_{t-1}$. It may be useful to superimpose these hyperbolas on the plot.

It may also be helpful to look at this plot as a pictorial representation of a two-by-two table. If $\{p_t\}$ were constant, the maximum likelihood estimate of ψ would be

$$\hat{\psi} = \frac{(\#\text{ of points in quadrant I})\ (\#\text{ of points in quadrant III})}{(\#\text{ of points in quadrant II})\ (\#\text{ of points in quadrant IV})} .$$

(The points would also appear in the same plotting position.) If the marginal probabilities do not vary a great deal, this relation suggests that counting points in each coordinate, or just obtaining some visual impression of the counts, may give information about ψ.

In summary, residual plots for the logistic regression model do not give the strong visual indication of serial dependence that they give in ordinary least squares. Information about serial dependence can be obtained through inspection of these plots, but the most striking features of the plots are not those that are most useful in detecting serial dependence. Experience or careful study is needed in order to extract the desired information. The score statistic is a much better indicator of serial dependence.

Chapter 8

Application to EKG Data

The inspiration for the serial dependence model came from work on the automatic classification of heart beats by their EKG traces. In this chapter I give a brief description of the problem. I also apply some of the procedures developed in this paper and I mention some of the difficulties that arise.

8.1 Background

In this section I give a brief description of the automatic beat classification problem. More details are given by Ngwengwe (1984).

Examination of EKG traces can give valuable information about the likelihood of future heart problems. Often life-threatening heart trouble such as ventricular fibrillation is preceeded by milder arrhythmia. Detection of abnormalities can therefore aid the physician in deciding whether preventative measures are required.

Figure 8.1 shows a typical normal beat. The curve is the electrical potential measured between two electrodes placed on the patient's chest. Each beat consists of a small P wave, a larger QRS complex, and a small T wave. A physician can detect abnormal beats such as premature ventricular

Figure 8.1. Idealized Normal Beat. This figure shows the components of a normal beat observed without noise on a single channel.

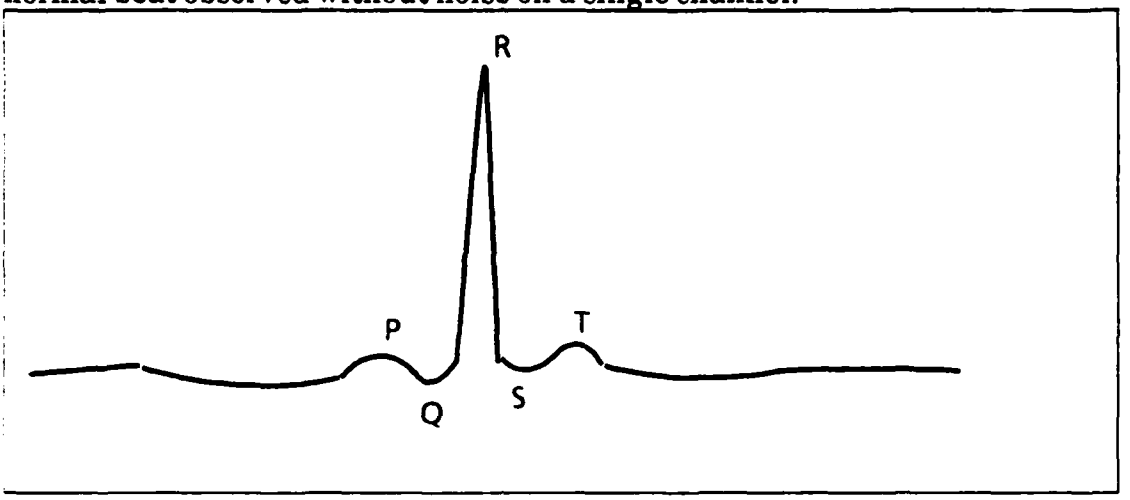
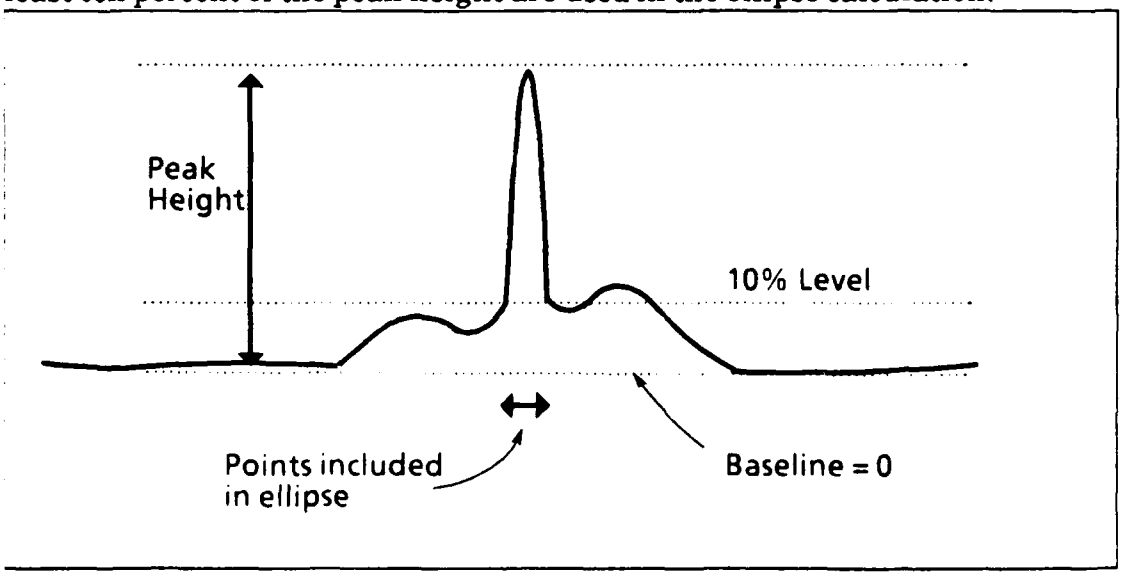


Figure 8.2. Choice of Points for Ellipse. The solid curve is the magnitude of the signal. Only consecutive points around the peak with a magnitude at least ten percent of the peak height are used in the ellipse calculation.



contractions (PVC's) by examining the EKG trace for beats that differ in some way from this normal form. Ngwengwe studied the use of automatic procedures for detecting abnormalities by computer.

This study used the MIT/BIH database. The database contains EKG measurements on forty-eight patients, each about thirty minutes in length. Each observation consists of a pair of measurements, giving the electrical potential between two pairs of electrodes placed on the chest along roughly perpendicular axes. Along with each trace is a set of beat locations and classifications provided by a cardiologist, so the number of beats and the type of each beat can be considered known for the purpose of this work.

Ngwengwe carried out a study of the ability of various features of the EKG traces to discriminate between normal heart beats and premature ventricular contractions. Among the techniques used to measure the power of each feature as a discriminating variable were linear discriminant analysis, recursive partitioning, and logistic regression.

Some of the best features were suggested by a simple graphical procedure. If the two components (or channels) of the EKG measurement are plotted against each other and observed over time, they appear to trace out an ellipse. The appearance of the ellipse is different for PVC's than for normal beats. Ngwengwe found that features associated with this ellipse provided excellent discrimination between normal beats and PVC's.

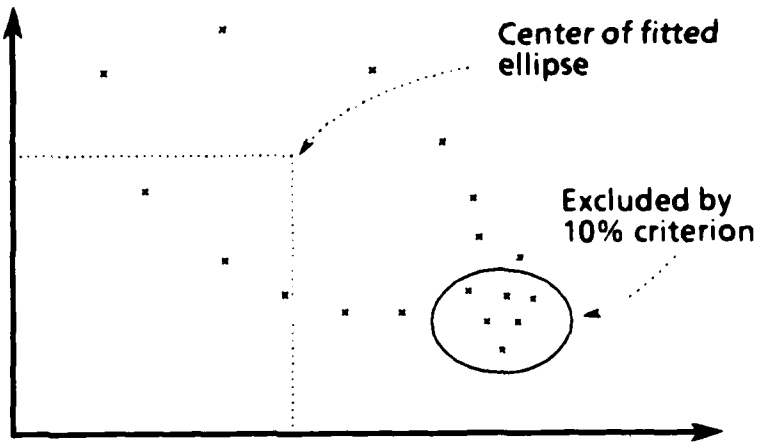
Figures 8.2 and 8.3 illustrate the procedure Ngwengwe used to obtain the ellipse features. Between each pair of beats is a relatively long stretch over which the signal is roughly constant. Taking the median signal over a long range therefore provides a "baseline" signal that can be subtracted from the entire range. This can be done for both channels. This provides an origin, and the distance of the two dimensional measurement from this origin is plotted in Figure 8.2.

The bulk of the apparent ellipse consists of points in the time range during which the magnitude of the signal is at least ten percent of its peak value. This is the cutoff line in the figure. Only points in this range are used in the ellipse calculation. For normal beats, this range generally includes only points from the QRS complex. For PVC's, on the other hand, this criterion may cause points from the P wave or the T wave to be included.

Figure 8.3 is a typical plot of the two components of the EKG trace. One end of the ellipse contains many points, including those along the baseline. These are excluded by the cutoff. The other points trace out the ellipse, and the distance between these points generally increases as the points move toward the opposite end of the ellipse.

The parameters of the ellipse can be estimated by computing the mean vector and covariance matrix for the points outside the cutoff. Because of the closer spacing at one end of ellipse, it is necessary to weight the points according to a scheme described by Ngwengwe. The ellipse then con-

Figure 8.3. Ellipse Parameters. The parameters of the ellipse are estimated by a method described by Ngwengwe (1983). Parameters used in this chapter are the number of points in the ellipse and the coordinates of the center.



sists of those points with a Mahalanobis distance from the center equal to two.

Ngwengwe investigated the ability of the five ellipse parameters and the number of points in the ellipse to discriminate between normal beats and PVC's. The three measurements he found most useful are the following:

NPOINTS: the number of points in the ellipse, a measure of the width of the beat.

XCENTER: the X coordinate of the center of the ellipse, a measure of the height along channel 1.

YCENTER: the Y coordinate of the center of the ellipse, a measure of the height along channel 2.

These are the features that I will use in this chapter.

8.1 Logistic regression

Various difficulties occur in trying to fit the probability of a PVC as a function of the above features by logistic regression. For many patients it is possible to separate the normal beats and the PVC's by a hyperplane in the three dimensional feature space. In some cases a single feature, usually NPOINTS, would separate the two beat types. This is a sign that the features work well, but it prevents the fitting of a logistic regression. I will refer to this as "perfect separation."

One assumption made in logistic regression is that observations are independent given their covariates. This does not seem to be a reasonable

assumption here, and that is the subject of this section. I will apply the techniques used in this paper to two of the patients.

The difficulties in fitting a serial dependence model exceed those in fitting an ordinary logistic regression. Clearly for those patients with perfect separation, the serial dependence model cannot be fit. Even without perfect separation, a high degree of serial dependence may lead to an infinite odds ratio estimate.

Another difficulty is a consequence of the excellent discrimination provided by these features. In many cases the fitted probability of a PVC is near zero for normal beats and near one for PVC's. When the odds ratio is large and two identical beat types appear in a row, the marginal probability of the observed pair may be very close to one. This leads to numerical problems in the maximum likelihood computations.

Logistic regression for patient 217 produces the following estimates:

	Coefficient Estimate	Standard Error
Intercept	3.946	1.242
NPOINTS	-.1228	.03121
XCENTER	-.0003469	.003575
YCENTER	-.1009	.01487

The score statistic is $W^{1/2} = 0.23419$, giving a one-step odds ratio estimate of 1.4850. There is perfect association, so the maximum likelihood estimate of the odds ratio is infinite.

To examine the effect on the coefficient estimates, I computed restricted maximum likelihood estimates of the coefficients with the odds ratio set at various values. The results are as follows:

Odds Ratio	Intercept	NPOINTS	XCENTER	YCENTER
2	3.911	-.122	-.00033	-.1000
4	3.879	-.121	-.00032	-.0979
8	3.898	-.122	-.00033	-.0941
16	3.963	-.125	-.00034	-.0883
32	3.922	-.125	-.00029	-.0825
64	3.810	-.122	-.00023	-.0778
128	3.677	-.119	-.00018	-.0745
256	3.573	-.116	-.00018	-.0723
512	3.490	-.114	-.00020	-.0709

Throughout this range the estimates remain within one half of a standard error of the logistic estimates, so the serial dependence does not seem to have had any adverse consequence on the coefficient estimates.

Patient 210 is an example of a data set with perfect separation. However if NPOINTS is not used the perfect separation disappears, so I will perform logistic regression using only the other two features. I do this in order to illustrate the effect of serial dependence, but of course NPOINTS is the best feature in this case and it should not be ignored in a search for the best model.

The results are as follows:

	Logistic Coef. Estimates	Logistic Std Errors	Serial Dep. Model Coef. Estimates
Intercept	-.315	.272	-.313
XCENTER	.0318	.0023	.0322
YCENTER	-.0288	.0039	-.0276

The score statistic is $W^{1/2} = 1.2979$, and the one-step estimate of the odds ratio is 4.67. The maximum likelihood estimate of the odds ratio is 1.79. Here again there is not a significant difference between the logistic and maximum likelihood estimates, as measured by comparison with the logistic standard errors.

Chapter 9

Summary

In this chapter I give a brief summary of the paper. I also comment on alternative models for serial dependence in binary regression and I present some generalizations of the serial dependence model.

9.1 Summary

In this paper I have proposed a regression model for binary time series, by analogy with the first order autoregressive model for normal time series. Dependence between successive observations is measured by the odds ratio, and this odds ratio is assumed constant over time. The process has a Markov property, so two observations are independent given an intervening observation. The marginal probability of $\{Y_t=1\}$ is a logistic function of covariates. In the special case of independence, the odds ratio is equal to one, and the model is equivalent to the ordinary logistic model.

With this model calculations of quantities involving marginal probabilities are quite simple. Calculations of joint probabilities are more complicated, but they are conventiently done by defining the quantity $\alpha_t = P[Y_t=Y_{t-1}=1]$, which is the solution of a quadratic equation. The parameters of a log linear representation for joint probabilities can be related to the parameters of this model, and the interactions terms are

simple functions of the odds ratio. A simple expression gives the odds ratio between observations that are not adjacent. A crude bound on the expression proves that the model generates ϕ-mixing sequences.

If a logistic model is fit to a process generated by the serial dependence model, inferences about the coefficients are suspect, because the standard errors of the estimates are not correct. However the closest logistic model to any serial dependence model is the one with the same coefficients if the distance is the Kullback-Leibler distance using the serial dependence model as the true model. This suggests that the ordinary logistic coefficient estimates ought to be consistent.

In fact the maximum likelihood estimates of the coefficients and the odds ratio are consistent under certain conditions, and a simulation shows no significant difference between the logistic and the maximum likelihood coefficient estimates. The maximum likelihood estimates can be calculated by Fisher's scoring method, which is equivalent to Newton's method with the second derivative replaced by its expectation. A simple estimator for the odds ratio is obtained by computing the ordinary logistic coefficient estimates and the score statistic for independence, and by finding the odds ratio for which the expected value of that statisic is the value actually observed. This estimate performs about as well as the maximum likelihood for moderate values of the odds ratio. For more extreme values is underestimates the magnitude of the log odds ratio, but it avoids the problem of infinite estimates.

The score statistic for testing independence is simply the autocovariance or its square root. The size and power of the test are well approximated by the values given by large sample theory. The accuracy of these approximations decreases as the odds ratio leaves the range [0.4, 2.5]. A higher order approximation does not improve the accuracy outside this range. For any given set of covariates an empirical approximation to the power function can be obtained, but it is not valid for other covariates. This suggests that the power depends on more than just the information and the odds ratio.

If observations are missing, the contribution to the score statistic (for testing independence) given by the dependence between observations on either side of the gap is infinitesimal, so the terms in the statistic must be summed over consecutive observations. An asymptotic relative efficiency calculation shows that under some circumstances a statistician may prefer to take a fixed number of observations spread out rather than consecutively. However these circumstances are not likely to occur in practical calculations, since they would require an extreme prior estimate of the odds ratio and they might require advance knowledge of the marginal probabilities.

Graphical displays of the residuals do not give vivid demonstration of serial correlation here as they do in least squares. However careful study of certain plots can be revealing.

Fitting of this model to features obtained from the MIT/BIH database of

EKG traces is complicated by perfect or near perfect separation of the features of normal beats and premature ventricular contractions. In one case with perfect association, restricted maximum likelihood estimates of the coefficients do not vary a great deal as the odds ratio is set at various values. In another case maximum likelihood estimation is possible and the estimated odds ratio is 1.79. The alternative estimator gives a higher value of 4.67.

9.2 Other models

Other models for serial dependence are possible. A direct generalization of the least squares model with an autoregressive error term is the following:

$$\log\,(p_t/(1-p_t)) = X_t'\beta + \varepsilon_t, \qquad \varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

where $\{u_t\}$ are indepenent normal random variables with mean zero and common unknown variance. However this is no longer a logistic regression model even when $\rho=0$. Under this model the sequence $\{p_t\}$ has a joint logistic-normal distribution. Some properties of this distribution are given by Aitchison and Shen (1980). Simple expressions for the moments are not possible, and maximum likelihood estimation of this model is likely to be quite difficult.

A simpler model is the one used by Korn and Whittemore (1979); Y_{t-1} is included as an explanatory variable for Y_t. I will refer to this model as the "lagged dependent variable" model. Under this model the conditional rather than the marginal probabilities take the logistic form. Fitting the model is quite simple, since it can be done by ordinary logistic

regression. Calculation of conditional probabilities is also simple.

As in the serial dependence model, the odds ratio between consecutive observations is constant. Since the process has a Markov property many of the results from Chapter 2 apply to this model as well. In particular the expression for the odds ratio between distant observations applies, and this model also generates *-mixing sequences. It is interesting to note that the expressions for ψ_{13} in equations [2.4] and [2.5] involve the quantity $u_{2(1)}$, a parameter in the log linear representation for the joint probabilities of three observations. These joint probabilities cannot be determined under the lagged dependent variable model, because the marginal probability of the first observation is unspecified. However the conditional probabilities are sufficient to calculate $u_{2(1)}$.

Any proof of the consistency and asymptotic normality of maximum likelihood estimates that is valid for longitudinal data is not likely to apply under the conditions assumed in this paper. However a proof of consistency along the lines of the one in Chapter 4 may be possible given the *-mixing property.

Calculation of marginal probabilities is difficult under the lagged dependent variable model. In particular the marginal distribution of Y_1 is not determined by the conditions I have stated; it requires a specified prior, or an assumed value or prior for Y_0. Then the marginal distribution at each time is determined. Unfortunately calculating the marginal distribution at time t requires summing over all possible values of Y_s, $s<t$.

9.3 Generalizations

The serial dependence model consists of two components: the relationship between the covariates and the marginal probabilities, and the serial dependence. Both of these could be generalized.

The logistic model is perhaps the most common binary regression model, but other models are possible. Most take the form $p_t = F(X_t'\beta)$, where $F(.)$ is some continuous cumulative distribution function. This is the most general model for which p_t is a continuous monotone function of $X_t'\beta$ and for which p_t approaches zero or one as $X_t'\beta$ approaches plus or minus infinity. Common choices for F, aside from the logistic, are the normal, extreme value, and uniform.

The analysis here could be repeated for these models. Some of the results, such as the ϕ^*-mixing property, depend only on the dependence and not on the form of F. These results apply for any choice of F.

The odds ratio does not extend to higher dimensions. But from Chapter 2, the constant odds ratio condition can be restated as follows: for all t,

$$\log P[Y_{t-1}=i,Y_t=j] = u(t) + (-1)^{i+1}u_1(t) + (-1)^{j+1}u_2(t) + (-1)^{i+j}u_{12},$$

where u_{12} is not a function of t. This suggests the following second order model, with the plus or minus sign taken as appropriate to the usual log linear model convention:

$$\log P[Y_{t-2}=i,Y_{t-1}=j,Y_t=k] = u(t) \pm u_1(t) \pm u_2(t) \pm u_3(t) \pm u_{12} \pm u_{23} \pm u_{13} \pm u_{123}.$$

Here the final four terms are not functions of time. It may make sense to require $u_{12}=u_{23}$, and the special case $u_{123}=0$ may also be interesting. Higher order generalizations can be defined similarly.

Bibliography

Aitchison, J., and S.M. Shen (1980), "Logistic-normal distributions: some properties and uses," Biometrika 67, 261-272.

Billingsley, P. (1961), Statistical Inference for Markov Processes, University of Chicago Press: Chicago.

Blum, J.R., D.L. Hanson, and L. Koopmans (1963), "On the strong law of large numbers for a class of stochastic processes," Z. Wahrsch. Verw. Gebeite 2, 1-11.

Chernoff, H. (1972), Sequential Analysis and Optimal Design, SIAM: Philadelphia.

Cox, D.R., and D.V. Hinkley (1974), Theoretical Statistics, Chapman and Hall: London.

Dunsmuir, W. (1981) "Estimation for stationary time series when data are irregularly spaced or missing," in Applied Time Series Analysis II, D. Findley, ed., Academic Press: New York.

Durbin, J., and G.S. Watson (1950), "Testing for serial correlation in least squares regression I," Biometrika 37, 409-438.

Durbin, J., and G.S. Watson (1951), "Testing for serial correlation in least squares regression II," Biometrika 38, 159-178.

Durbin, J., and G.S. Watson (1971), "Testing for serial correlation in least squares regression III," Biometrika 58, 1-19.

Feinberg, S. (1981), The Analysis of Cross-Classified Categorical Data, second edition, MIT Press: Cambridge, MA.

Hall, P., and C.C. Heyde (1980), Martingale Limit Theory and Its Application, Academic Press: New York.

Harris, P., and H.W. Peers (1980), "The local power of the efficient scores test statistic," Biometrika 67, 525-529.

Honohan, P., and C. McCarthy (1982), "On the use of Durbin-Watson type statistics when there are missing observations," The Statistician 31, 149-152.

Kedem, B. (1980), Binary Time Series, Marcel Dekker: New York.

Keenan, D.M. (1982), "A time series analysis of binary data," J. Amer. Statist. Assoc. 77, 816-821.

Korn, E.L., and A.S. Whittemore (1979), "Methods for analyzing panel studies of acute health effects of air pollution," Biometrics 35, 795-802.

Ngwengwe, A.M. (1981), ECG Analysis: Automatic Classification of Heart Beats based on ECG traces, S.M. thesis, M.I.T.

Peers, H.W. (1971), "Likelihood ratio and associated test criteria," Biometrika 58, 577-587.

Rao, C.R. (1973), Linear Statistical Inference and Its Application, Wiley: New York.

Savin, N.E., and K.J. White (1978), "Testing for autocorrelation with missing observations," Econometrica 46, 59-67.

Wald, A. (1949), "Note on the consistency of the maximum likelihood estimate," Ann. Math. Statist. 20, 595-601.

Zeger, S.L., K.Y. Liang, and S.G. Self (1985), "The analysis of binary longitudinal data with time-independent covariates," Biometrika 72, 31-38.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 38	2. GOVT ACCESSION NO.: AD-A161 274	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): A Model for Serial Dependence in Logistic Regression		5. TYPE OF REPORT & PERIOD COVERED: Technical Report
		6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Thomas P. Lane		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0555
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Statistics Center, Massachusetts Institute of Technology, Cambridge, Mass. 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-042-331)
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Statistics & Probability Code 436, Arlington, Va. 22217		12. REPORT DATE: September 1985
		13. NUMBER OF PAGES:
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

see reverse side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A model is proposed for binary time series with marginal probabilities given by logistic regression on explanatory variables, by analogy with the first order autoregressive error model for least squares regression. Measurements at adjacent time points are assumed to have an odds ratio that is not equal to one and that is constant as a function of time. Measurements separated in time are assumed to be conditionally independent given an intervening observation.

Consequences of using an ordinary logistic model in the presence of serial dependence are explored. The closest logistic model, defined as the one with the minimum Kullback-Leibler distance, is shown to be the one with the same marginal probabilities. Consistency of the maximum likelihood estimator of the serial dependence model is proved under certain conditions, and a procedure for finding these estimates is given.

Properties of the model are found, including expressions for the joint probabilities and the odds ratio between observations separated in time. The model is shown to generate*-mixing processes.

A score test is derived in order to test for independence after performing an ordinary logistic regression, and properties of this test are explored. The effects of missing data on the score test and on estimation of the odds ratio (with known coefficients) are presented.

The model is applied to the problem of automatic classification of EKG data based on feature extraction. A positive serial dependence is found in the examples presented.

END

FILMED

1-86

DTIC